EGGNet: Learning Temporal Boundaries of Glottal Cycles from EGG

Rui Feng, *Graduate Student Member, IEEE,* Yuang Chen, *Graduate Student Member, IEEE,*
Jun Du, *Senior Member, IEEE,* Zhenhua Ling, *Senior Member, IEEE,* Jiahong Yuan, *Member, IEEE*

***Abstract*—The electroglottograph (EGG) is extensively adopted as a direct and reliable physiological signal for high-quality pitch annotation. However, popular methods represented by *peakdet* suffer from false alarms and missed detections in glottal closure instant (GCI) detection. This paper proposes EGGNet, an innovative wide-context multi-branch convolutional model, which overcomes these bottlenecks through three key technical breakthroughs. Firstly, EGGNet incorporates a temporal evaluation optimization with a tolerance-window-based GCI detection algorithm, which mitigates temporal localization sparsity-induced distortions and limits matching errors to ± 20 sampling points amid tens of thousands of samples per second. Secondly, EGGNet proposes a greedy-based one-to-one matching strategy to eliminate the multi-to-one or one-to-multi evaluation bias common in traditional methods, ensuring precise alignment with ground-truth labels. Furthermore, a cascaded supervision architecture is introduced to foster a collaborative training mechanism between soft and hard labels, effectively balancing sensitivity and localization accuracy. Extensive experiments trained on the PTDB-TUG dataset demonstrate that EGGNet significantly outperforms *peakdet* and achieves F1 scores of 96.6%, 96.2%, and 90.3% on CSTR-FDA, KEELE, and MOCHA-TIMIT, respectively. Comprehensive ablation studies further confirm that EGG exhibits superior performance on high-noise corpora, with soft-hard cascaded supervision improving the precision–recall balance by 6.5% over single-supervision schemes.**

***Index Terms*—Electroglottography (EGG); Glottal Closure Instant (GCI); Deep Learning; Binary Localization**

I. Introduction

As one of the foundational mediums for human communications, pitch has long been extensively used in various speech processing tasks, such as speech synthesis, speech recognition, and affective computing [1]–[3]. The successful execution of these tasks heavily relies on reliable extraction methods of acoustic characteristics [1], [2], [4], among which the fundamental frequency (F0) serves as a critical foundational characteristic [5], [6]. In recent years, the speech signal-based F0 extraction algorithms (e.g., YIN [7], CREPE [8], and pYIN [9]) have made remarkable advancements. However, regardless of how advanced these F0 extraction algorithms become, their performance is essentially constrained by the quality of the labeled data exploited for training and evaluation [8], [10]–[12].

Currently, the vast majority of F0-labeled datasets (e.g., [13]–[17]) are generated from either manual labeling or automatic speech algorithms, which have been demonstrated to introduce labeling errors or lack robustness in diverse acoustic conditions. The technical challenge has prompted researchers to consider leveraging a more direct and reliable physiological signal, i.e., electroglottography (EGG), as the basis for high-quality speech feature annotation [6], [18]. The EGG signal directly measures fluctuations in electrical impedance caused by changes in vocal fold contact area, thereby accurately reflecting the glottal opening and closing movements [6], [19]. Compared to methods that indirectly infer these dynamics through acoustic signals, the EGG signal enables more precise and stable determination of the onset and offset of speech cycles [18], [19]. These advantages make EGG signals an ideal annotation source for constructing high-precision speech datasets [20]. Unfortunately, due to the relatively high cost and operational complexity of EGG signal acquisition devices, such high-quality annotated data remains relatively scarce. Concurrently, existing methods for annotating cycle boundaries based on EGG signals still have significant room for optimization, particularly in improving temporal resolution and adaptability to pathological voices [21]–[24].

Many existing studies have shown that the positive peaks of the differential electroglottography (dEGG) signal exhibit a high degree of consistency with glottal closure instants (GCI)[1], making these peaks a reliable marker for speech cycle boundaries [26]. The classic peak detection algorithm, i.e., *peakdet* [23], [27], has been widely applied in this context [21]–[24]. However, preliminary experiments on real-world datasets indicate that this traditional peak detection method still suffers from noticeable false positives and misses under complex signal conditions, achieving a precision of approximately 48.9%, a recall of about 84.3%, and an F1 score of around 61.9% [22][2]. Such unsatisfactory detection

Rui Feng and *Yuang Chen* are co-first authors with equal contributions, and the sequence of authorship does not reflect the magnitude of their contributions. This work is supported in part by the National Social Science Foundation of China (23AYY012) and USTC (YD2110002303). Rui Feng, Yuang Chen, Jun Du, Zhenhua Ling, and Jiahong Yuan are with the National Engineering Research Center of Speech and Language Information Processing, University of Science and Technology of China (USTC), Hefei, P. R. China. Email: {fengruimse, yuangchen21}@mail.ustc.edu.cn, {jundu, zhling, jiahongyuan}@ustc.edu.cn. The corresponding author is Prof. *Jiahong Yuan.*

[1]The GCI refers to the moment during a complete vocal cord vibration cycle when the vocal cords transition from an open state to a fully closed state, which is a critical event with clear physiological significance in the process of speech production [25]. At the signal level, GCI corresponds to a rapid change in glottal impedance. dEGG signals typically exhibit prominent local extrema near the time of glottal closure, and the dEGG waveform obtained by differentiating the EGG signal can accentuate abrupt transitions corresponding to vocal fold contact. Characteristic peaks, either positive or negative, often align with points of maximum slope or rapid decline in the EGG waveform and are widely used as reliable indicators for locating GCIs.

[2]**http://voiceresearch.free.fr/egg/softwares.htm**

performance clearly falls short of the stringent requirements for high-quality speech data annotation.

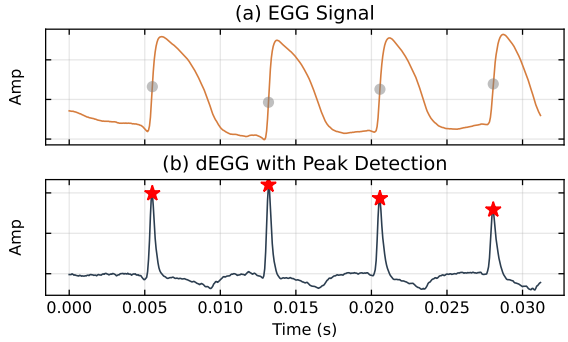

Fig. 1: Peak detection algorithm annotation under ideal conditions.

As illustrated in Fig. 1, we show the performance of using the *peakdet* algorithm for the GCI detection under ideal conditions. However, when the signal-to-noise ratio (SNR) of the EGG signal is low or the background noise is large during unvoiced portions/segments, the GCI detection that is directly annotated by *peakdet* has significant errors and omissions in the annotation of GCI detection, which is clearly illustrated in Fig. 2. In particular, the *peakdet* algorithm can identify GCIs by locating negative peaks in the dEGG waveform. However, *peakdet* heavily relies on the local extremum detection with fixed thresholding and does not incorporate temporal context or signal adaptation. Moreover, although *peakdet* is lightweight and straightforward to implement, it is also highly sensitive to noisy or ambiguous signals, which limits its robustness in real-world conditions.

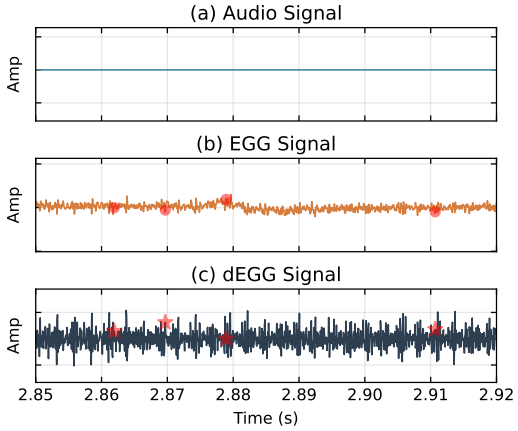

Fig. 2: The mislabeled segment of *peakdet* during unvoiced segments. (a) The original audio signal is in a non-acoustic state. (b) The EGG signal corresponds to the system's background noise. (c) The dEGG signal that is the differential of the background noise.

A. Motivations and Challenges

Combining Fig. 1 and Fig. 2, it can be seen that the so far widely used *peakdet* exhibits a serious lack of accuracy in F0 labeling, even though it is almost the most advanced EGG detection algorithm. In particular, *peakdet* has significant limitations in its performance under low SNR regions, unvoiced segments, and poor signal quality. These limitations have long been ignored and have not been well addressed. If they continue to be overlooked, it will negatively impact the accuracy of data analysis, manual workload, medical diagnostic risk, and technological advances. As a result, in order to achieve robust F0 extraction using EGG signals, it is necessary to propose effective solutions to tackle the limitations of the *peakdet* in terms of labeling, to realize the complementation of missed labels and the deletion or correction of mislabeling of the differential peak points of the DEGG. The desired algorithm performance for labeling EGG signals for robust F0 detection can be shown in Fig. 3 intuitively, as follows:

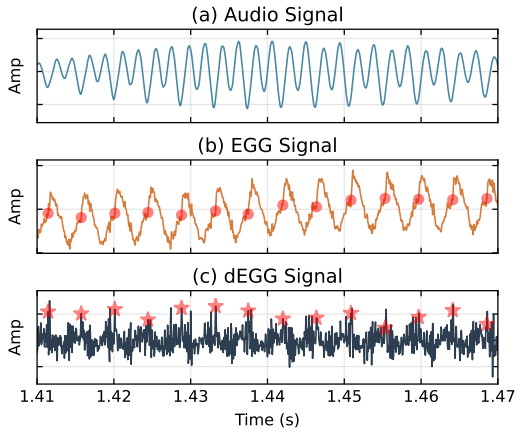

Fig. 3: The desired algorithmic performance for labeling temporal boundaries of Glottal Cycles of EGG signals on the premise of not missing any vocal fold vibration cycle.

Nevertheless, to fulfill the algorithmic performance illustrated in Fig. 3 is not a task that can be solved easily, and it requires verifying the annotation of the *peakdet* algorithm, including completing annotations and removing mislabeled differential peak points, to form a verified reference peak position. In addition, it also needs to calculate the Gaussian distance from the reference position as confidence in order to obtain more accurate label annotations. On this basis, it is necessary to propose corresponding algorithms to effectively segment the EGG signal waveform and accurately predict its glottal closure time, accurately characterize the deviation between the closure time and the confidence level of manual labeling, only in this way can noise interference be better eliminated and higher robustness be achieved.

B. Literature Review

F0 is a critical parameter in speech signal processing, playing a significant role in tasks such as speech synthesis, speech recognition, and affective computing [1], [2], [5]. Over the years, researchers have developed various F0 extraction algorithms, including the time-domain-based YIN [7], the probabilistic-based pYIN [9], and the deep learning-based CREPE [8]. These algorithms have made significant progress under ideal conditions, but their performance is highly dependent on the quality of the annotated data used for training and evaluation [10], [11], [28]. For example, Jong Wook Kim *et al.* in [8] proposed CREPE and detailed that it was trained in

datasets of synthesized audio to achieve perfect control over the F0 of the reusing signal for the purpose of guaranteeing a perfectly objective evaluation. If the ground truth dataset of F0 (e.g., pYIN) is inaccurate, the evaluation metrics (e.g., gross pitch error (GPE), voicing decision error (VDE)) may misrepresent the true capacities of algorithms, potentially introducing noise during training and leading to performance degradation. This was highlighted in comparative studies [10], which evaluated multiple F0 extraction algorithms but assumed the labeled data was accurate, potentially overlooking the sensitivity to data quality. The evidence leans toward a direct correlation between labeled data quality and algorithm performance. Additionally, the authors in [11] introduced the F0 frame error metric, which integrates the GPE and VDE, along with a GPE-VDE curve to analyze their tradeoff, and proposed a model-based U/V classification frontend that significantly reduces VDE and FFE under noisy scenarios when applied to various F0 tracking algorithms.

To address the issue of annotated data quality, researchers have turned to EGG as an alternative method for obtaining high-precision F0 measurements. EGG is a non-invasive technique that directly reflects vocal fold vibrations by measuring changes in laryngeal electrical conductance [19]. The positive peaks in the derivative of the EGG signal (dEGG) are highly correlated with the GCI and can be used for precise F0 calculation [20]. Therefore, EGG is typically used as the "gold standard" for evaluating F0 extraction algorithms [29]. For example, [29] pointed out that EGG recordings are typically used to provide reference values for F0 in speech corpora, although they are not always available in all corpora and may not always align perfectly with the pitch characteristics of acoustic signals in some cases. Traditionally, methods for extracting F0 from EGG signals have relied on peak detection algorithms, such as the *peakdet* algorithm, which determines GCI by detecting positive peaks in the dEGG signal [24]. This method performs well under normal speech conditions but is prone to false detections or missed detections under low signal-to-noise ratio (SNR) or unvoiced segments (for example, when background noise is high) [23]. Furthermore, the *peakdet* algorithm lacks sufficient temporal resolution and adaptability when processing pathological speech, which limits its application in complex speech scenarios.

In recent years, machine learning techniques have begun to be applied to EGG signal analysis to improve the accuracy and robustness of F0 extraction. For example, [30] explored the use of neural networks to estimate F0 and GCI from EGG signals, demonstrating the potential of data-driven methods in this field. Additionally, [31] provided a systematic review of EGG-based digital speech quality assessment systems, highlighting the application of machine learning in speech pathology diagnosis. These studies indicate that machine learning methods can overcome some limitations of traditional peak detection algorithms, but further optimization is needed to address the challenges of low-quality signals and pathological speech. Despite the progress made by the aforementioned methods, existing technologies still have significant room for improvement in terms of temporal resolution and adaptability to pathological speech. Especially in low SNR environments or unvoiced segments, traditional methods and initial machine learning approaches struggle to meet the stringent requirements for high-precision F0 annotation. This motivates the development of novel deep learning models, which aim to significantly improve the quality of F0 annotation datasets.

C. Main Contributions

In order to effectively overcome the aforementioned challenges of the state-of-the-art F0 extraction algorithms, presented by *peakdet*, this paper proposes an innovative deep learning model termed EGGNet for the first time, which consists of the proposed F0 extraction algorithm based on temporal neural network that combines with EGG signals as a high-precision benchmark, and and comprehensively optimizes the consistency of the peak annotation of EGG signals through manual proofreading, thus realizing the automated annotation process that is extremely efficient and accurate. In particular, the proposed EGGNet framework is dedicated to the automatic detection of periodic boundaries in EGG signals, with a well-designed multi-scale convolutional network architecture and a robust adaptive thresholding method. Compared with the state-of-the-art positive peak detection algorithm *peakdet*, EGGNet shows substantial improvement in key performance metrics such as Precision, Recall, and F1 score, and we discovered interesting patterns in the embeddings learned by our model. This improvement allows for more accurate annotation of speech cycle boundaries, resulting in the creation of higher-quality F0 annotated datasets. The primary contributions of this paper are summarized as follows:

- Towards the current lack of precise annotation data for GCI in EGG signals, this paper proposes and implements a human-machine collaborative GCI annotation framework, which combines manually drawn threshold trajectories by annotators with automated verification by subsequent computational methods, including consistency comparison between GCI-derived F0 and reference F0. This framework aims to generate internally consistent and physiologically plausible GCI annotations. Based on this, a high-quality GCI reference dataset covering multiple public corpora and spanning approximately **24** minutes has been successfully constructed, providing an important benchmark for research in this field. We have made the corpus open-source, and it can be downloaded from the link: **https://github.com/RuiFeng-USTC/CYA-EGGNet-data**.
- Next, this work proposes a novel deep learning network architecture named EGGNet, which is specifically designed for high-precision GCI detection from a single EGG signal. The core innovation of EGGNet lies in leveraging parallel multi-branch and multi-scale convolutional modules. By systematically employing one-dimensional convolutional layers with different-sized convolutional kernels across different branches or within the same branch, the architecture effectively captures and integrates the complex patterns associated with GCI events across various temporal scales within the EGG signal.
- The proposed EGGNet model designs and implements a novel two-stage supervised learning strategy based on

"smooth target preliminary approximation and precise target final discrimination". This strategy first guides the model to learn the "soft target" representation of the GCI in the time domain after Gaussian smoothing, and then refines the precise instantaneous position of the GCI (i.e., "hard target") based on this. In particular, the model achieves end-to-end GCI sequence prediction, with the output GCI position sequence strictly aligned in time with the input EGG signal, thereby significantly improving the accuracy of GCI localization and the efficiency of model training.

- Finally, through comprehensive experimental evaluations on multiple publicly available EGG corpora, the high effectiveness and superior performance of the EGGNet model in GCI detection tasks were systematically verified, with particular attention paid to its robustness in noisy environments. The experimental results show that even under conditions of low signal quality, EGGNet is still able to maintain significant performance advantages, effectively proving the advanced nature and practical value of the model.

The remainder of this paper is organized as follows. In Section. II, we first introduce the dataset description. In Section III, the methodology involved in this work is presented. In section IV, we introduce the proposed EGGNet framework in detail, followed by the extensive performance evaluation and comparisons. Finally, the conclusion and future outlook are given in Section VI.

II. Description and Pre-analysis of Corpus Datasets

In this section, we introduce the relevant datasets used in this paper, including CSTR-FDA [13], KEELE [32], PTDB-TUG [15], and MOCHA-TIMIT [16]. Notably, PTDB-TUG is used as the training dataset, while the other three datasets are used as the test datasets. The detailed descriptions of these datasets are introduced below.

A. CSTR-FDA

The CSTR-FDA dataset was first released by C. Bagshaw in [13], which has become an extensively used automatic method for evaluating the rhythm, pitch, stress, and intonation patterns of speech in the context of language learning. The CSTR-FDA dataset used for evaluation contains approximately five minutes of speech. The speech signals of CSTR-FDA were recorded using a close-talk microphone and an EGG with synchronous acquisition in an anechoic chamber. The CSTR-FDA dataset consists of 50 sentences, each read aloud by an adult male and an adult female, both of whom have non-pathological voices. This corpus contains a large number of voiced fricatives, nasals, liquids, and glides.

B. KEELE

The KEELE dataset was released by F. Plante*et al.* in [14], introducing a comprehensive dataset designed to evaluate and compare various pitch extraction algorithms in speech processing. The KEELE dataset provides a standardized set of speech and laryngograph data from 15 speakers reading phonetically balanced texts. The publicly available version includes data from 5 adult male and 5 adult female speakers. Recordings were performed using a head-mounted microphone and an EGG with synchronous acquisition. These speech signals were recorded through the digital audio tape (DAT) system and digitized at a sampling rate of 20 kHz with 16-bit resolution. Notably, this database has been extensively leveraged and significantly influenced the development and assessment of pitch detection methods across multiple applications, including objective evaluation and development of pitch extraction algorithms [33]–[37], improvement of speech recognition systems [38], advancement in voiced-unvoiced detection (VUD) [39], etc.

C. PTDB-TUG

The PTDB-TUG dataset was first released by Pirker *et al.* in [15], which focuses on the development of a pitch tracking corpus and the evaluation of pitch tracking algorithms in the context of multi-pitch tracking scenarios. The PTDB-TUG database consists of speech data from 10 male and 10 female native English speakers. A total of 2,342 sentences selected from the TIMIT corpus were used in the recordings. On average, each speaker read approximately 236 sentences, resulting in a total of 4,720 paired recordings of audio and EGG signals. The recordings were conducted in a studio environment using a head-mounted microphone and a laryngograph with a neckband electrode, both sampled synchronously at 48 kHz with 16-bit resolution.

D. MOCHA-TIMIT

Narayanan *et al.* firstly released the MOCHA-TIMIT dataset in [16], which introduces a unique multimodal dataset that combines real-time MRI with other speech-related measurements, such as acoustics and motion capture, to investigate the physiological and articulatory processes involved in speech production. This corpus also provides detailed insights into how articulatory movements, such as tongue and lip positioning, relate to speech sounds in real time. The sentences in the MOCHA-TIMIT corpus were taken from 460 sentences in the TIMIT corpus, mainly designed to cover all phonemes of American English and include phonetic connectivity phenomena in spoken English, such as assimilation, weakening, dropout, and merging. This corpus design facilitates systematic comparison with other studies that use the same sentence. The dataset contains 10 speakers, including 5 females and 5 males.

E. Pre-analysis of Corpus Datasets: Modeling and Evaluation from the Perspective of Signal Features

Next, we first perform some pre-analysis on the EGG signals in the aforementioned corpus datasets. By analyzing the signals of these corpus datasets, we found that the EGG signals in these datasets can exhibit some irregular fluctuations and background noise, which are mainly attributed to individual differences in glottal anatomy, electrode contact conditions, phonation patterns, and natural variations in the recording environment. Although these glitches are inevitable in the EGG signal acquisition process and have been reported in

several publicly available corpora, these corpora only provide limited documentation on specific hardware models and acoustic configurations, making precise quantitative comparisons impractical. To this end, this paper focuses on modeling and evaluating from the perspective of signal features.

To evaluate the quality of the EGG signals, this study defines the signal-to-noise ratio (SNR) for each speech sample, based on the energy contrast between voiced and silent segments. The detailed procedure can be formulated as follows:

Firstly, the raw EGG signal is processed using a zero-phase high-pass filter with a cutoff frequency of 50 Hz to eliminate low-frequency drift. Then, the signal is segmented into voiced and unvoiced regions based on a pre-annotated mask. The average power for each voiced or unvoiced region, i.e., P_{voiced} and P_{unvoiced}, can be calculated separately as follows:

$$P_{\text{voiced}} = \frac{1}{N_{\text{voiced}}} \sum_{t \in \mathcal{T}_{\text{voiced}}} x(t)^2, \tag{1a}$$

$$P_{\text{unvoiced}} = \frac{1}{N_{\text{unvoiced}}} \sum_{t \in \mathcal{T}_{\text{unvoiced}}} x(t)^2, \tag{1b}$$

where $x(t)$ indicates the high-pass filtered signal, N denotes the number of samples in the corresponding segment, and $\mathcal{T}_{\text{voiced}}$ and $\mathcal{T}_{\text{unvoiced}}$ represents the sets of voiced and unvoiced segments. Based on the definitions of (1a) and (1b), the SNR for each speech segment can be calculated as follows:

$$\text{SNR (dB)} = 10 \cdot \log_{10} \left(\frac{P_{\text{voiced}}}{P_{\text{unvoiced}}} - 1 \right). \tag{2}$$

The metric SNR (dB) reflects the energy dominance of the voiced segments relative to the unvoiced segments (i.e., background noise) within the same signals, capturing the contrast in signal strength without relying on absolute amplitude or speaker-specific characteristics. As shown in Fig. 4, we analyze the distribution characteristics of the four used corpus datasets and present them using violin plots, where all results were independently calculated based on each speech in the corpus database.

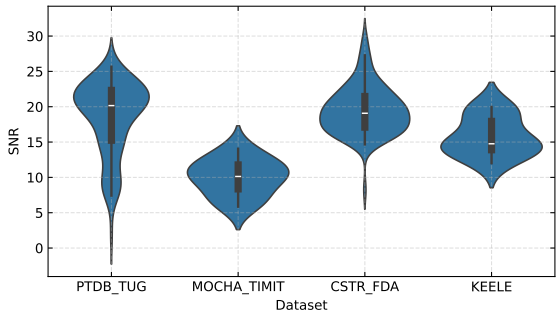

Fig. 4: The SNR distribution across the corpora datasets CSTR-FDA, KEELE, PTDB-TUG, and MOCHA-TIMIT.

Fig. 4 reveals the distribution of these four corpus datasets containing EGG signals, CSTR-FDA, KEELE, PTDB-TUG, and MOCHA-TIMIT, on the SNR metric. It can be observed that the SNR distribution range of PTDB-TUG is relatively wide, covering various situations from clean to heavily noisy, making it suitable as a training set to enhance the generalization ability of the model. On the other hand, CSTR-FDA, KEELE, and MOCHA-TIMIT represent test conditions with different levels of noise, including high, medium, and low, respectively, forming a complete test set suitable for comprehensively evaluating the robustness of EGG signal processing algorithms under various acoustic challenges.

III. The Methodology of Our Work

In this section, we introduce the methodology of this work, including the F0 extracted from GCI annotations, the conventional *peakdet* baseline, the GCI annotation procedure, label construction and enhancement, and a detailed introduction to the proposed EGGNet model framework.

A. The Extracted F0 from GCI Annotations

As illustrated in Fig. 5, we present the F0 extracted from GCI labeling corresponding to these four corpus datasets, as well as the erroneous labeling of EGG signals for GCI candidates calculated by *peakdet*. Fig. 5 (a), (b), (c), (d) illustrate the ground truth (GT) of F0, F0 calculated by manually annotated GCI, and F0 calculated by *peakdet* annotated GCI. It can be observed that on the corpus datasets PTDB-TUG, CSTR-FDA, KEELE, and MOCHA-TIMIT, there is an extremely close relationship between Manual GCI F0 and GT F0, with the two curves almost overlapping, which indicates that Manual GCI F0 generates highly reliable F0 trajectories and fully demonstrates the accuracy and effectiveness of our proposed Manual GCI F0. On the other hand, it can also be seen that there are significant differences between *peakdet* GCI F0 and GT F0 in the corpus datasets PTDB-TUG, CSTR-FDA, KEELE, and MOCHA-TIMIT, which are especially significant in PTDB-TUG, KEELE, and MOCHA-TIMIT. Therefore, Manual GCI F0 clearly achieves more accurate F0 extraction performance.

Furthermore, Fig. 5 (e), (f), (g), (h) shows the relationship between the EGG annotation obtained using *peakdet* GCI F0 and the actual EGG signals. It can be observed that in the PTDB-TUG, CSTR-FDA, KEELE, and MOCHA-TIMIT corpus datasets, the EGG annotations obtained using *peakdet* GCI F0 exposed the areas where *peakdet* algorithm incorrectly identified GCI in the region where the two F0 curves diverged. In particular, typical errors include missed detections in the low-amplitude vocalization process of the KEELE corpus, as shown in Fig. 5 (g), and incorrect detection of background noise in the silent tail of MOCHA-TIMIT vocalization, as shown in Fig. 5 (h).

B. GCI Annotation Procedure

In this study, GCI annotations were obtained using a semi-automatic, interactive procedure based on dEGG waveforms. For each continuous voiced segment, annotators manually drew a reference curve to approximately delimit the region where glottal closure events were likely to occur, as shown in Fig. 6. The algorithm then automatically selected the local maximum within each region above the curve as the GCI for that segment. This approach was adopted out of practical necessity, particularly in cases where low SNRs made it difficult for

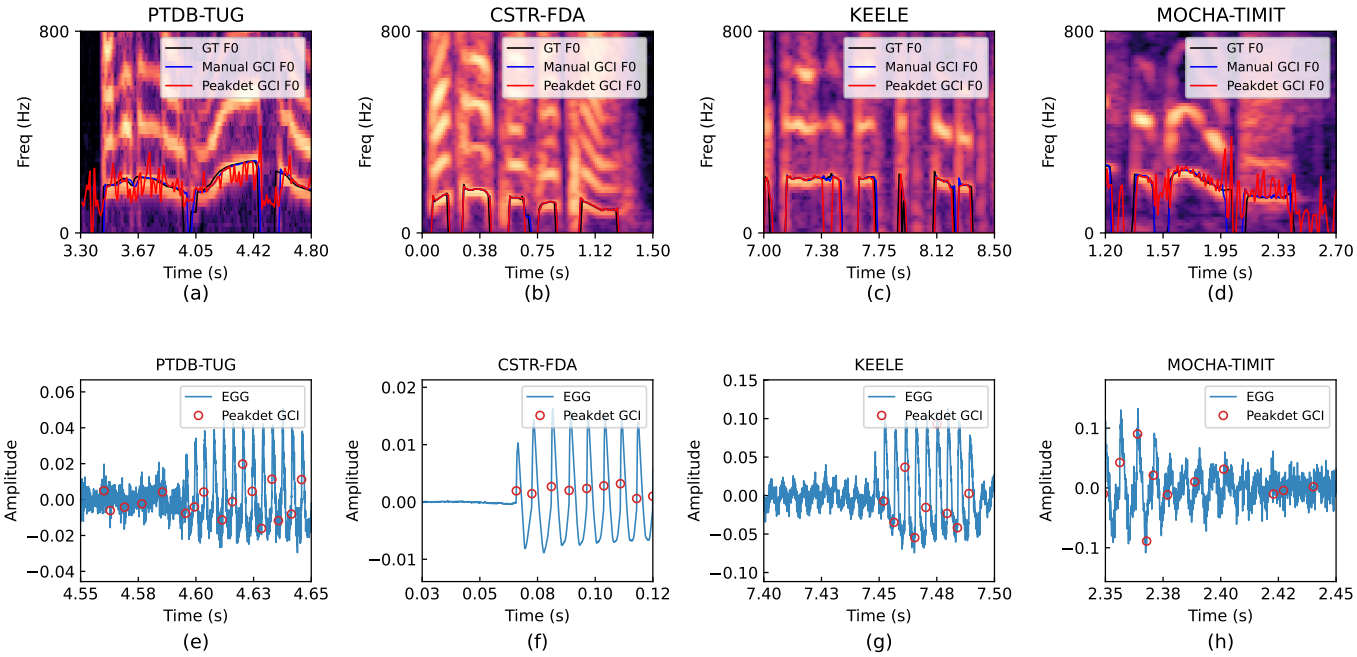

Fig. 5: The ground truth (GT) F0, F0 calculated by manually annotated GCI, and F0 calculated by *peakdet* annotated GCI on the corpus datasets PTDB-TUG (a), CSTR-FDA (b), KEELE (c), and MOCHA-TIMIT (d). And the relationship between the EGG annotation obtained using *peakdet* GCI F0 and the actual EGG signals on the corpus datasets PTDB-TUG (e), CSTR-FDA (f), KEELE (g), and MOCHA-TIMIT (h).

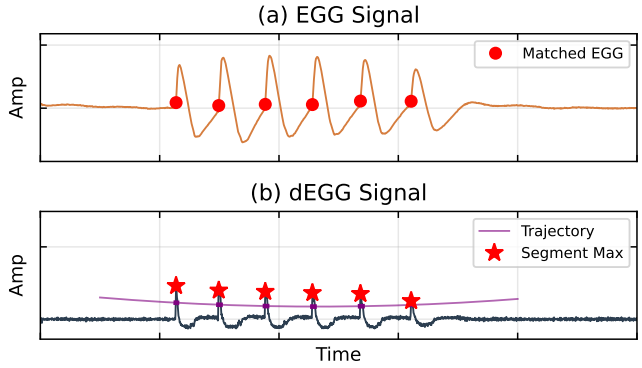

Fig. 6: The manual annotation of GCI.

fully automatic methods to reliably detect GCIs. The use of an interactive procedure improved the completeness and consistency of the annotations. Throughout the entire dataset, GCIs were labeled for all visually identifiable glottal cycles, without omitting any segments.

Additionally, within the training dataset, particularly the PTDB-TUG corpus, we have additionally recorded signal status of the location of each GCI, such as whether it is situated in the area of extremely low amplitude, or whether the dEGG peak at that location is morphologically blurred and not sharp enough. These additional attributes serve solely as prompt information during the training phase to assist the model in better identifying challenging samples. They are not utilized during the testing phase and do not affect the evaluation labels, ensuring the objectivity of the model assessment. In order to validate the effectiveness of our annotations, we have employed an indirect evaluation strategy to demonstrate their reliability. On the one hand, we will retain some of the annotated results as potential open-source materials for future use; on the other hand, we will compare the F0 derived from the GCI annotations with the reference F0 in the dataset, and contrast these results with those obtained from *peakdet*) for the same task, thereby indirectly illustrating the rationality and consistency of our GCI annotations, as illustrated in Table I.

TABLE I: Comparison of MAE and RPA using manual GCI annotation(M) vs. *peakdet*-generated GCI(*P*). Units: MAE in Hz, RPA in %.

Dataset	MAE (M)	MAE (*P*)	RPA (M)	RPA (*P*)
PTDB_TUG	10.07 Hz	30.92 Hz	83.94 %	68.41 %
MOCHA_TIMIT	10.26 Hz	76.50 Hz	83.62 %	30.60 %
CSTR_FDA	6.92 Hz	9.21 Hz	91.56 %	89.71 %
KEELE	7.48 Hz	20.00 Hz	88.22 %	78.57 %

The task of GCI detection is fundamentally a sparse temporal event localization problem, where the model must identify a very small number of true GCI events from tens of thousands of samples per second. This inherently leads to a highly imbalanced classification scenario. In particular, the number of positive instances is extremely small compared to the overwhelming number of background (negative) samples. Direct point-by-point evaluation under such an imbalance would bias the results heavily toward the background class, leading to a distorted assessment of detector performance. To effectively address this issue, as described by **Algorithm 1**, this paper proposes the tolerance window-based evaluation strategy, wherein a predicted GCI is considered a valid hit (i.e., true positive, TP) if the distance between it and a true GCI does not exceed 20 sampling points (i.e., $|p - g| \leq \delta$), where p and g indicates the predicted sequence and ground

true, and δ denotes the tolerance. In this paper, the tolerance δ is set to 20 sampling points, which approximately equals to $1.25ms$ at the sampling rate of 16 kHz, referencing the evaluation practices in the field of ECG QRS detection, where a detector temporal tolerance (DTT) of 60 to 150 *ms* is commonly used for metric calculations [40]. Meanwhile, in order to prevent one-to-many or many-to-one matching errors during evaluation, this paper has also implemented a greedy-based one-to-one matching strategy. Each ground-truth GCI serves as the reference, and the algorithm iterates through all reference points to find the closest predicted point within the tolerance window that has not been assigned to any other reference. This strategy enforces two constraints:

- Each predicted point is only allowed to be matched at most once.
- Each ground-truth GCI is only allowed to be matched by at most one predicted point.

Algorithm 1 Tolerance-Window Based GCI Detection Algorithm

Require: Predicted sequence $p(t)$, ground-truth $g(t)$, tolerance δ
Ensure: True Positive (TP), False Positive (FP), False Negative (FN)

$P \leftarrow \{t \mid p(t) = 1\}$
$G \leftarrow \{t \mid g(t) = 1\}$
$\mathcal{M} \leftarrow \emptyset, \quad TP \leftarrow 0$
for each $g \in G$ **do**
  $U \leftarrow P \setminus \mathcal{M}$
  $C \leftarrow \{p \in U \mid |p - g| \leq \delta\}$
  if $C \neq \emptyset$ **then**
    $p^* \leftarrow \arg\min_{p \in C} |p - g|$
    $\mathcal{M} \leftarrow \mathcal{M} \cup \{p^*\}$
    $TP \leftarrow TP + 1$
  end if
end for
$FP \leftarrow |P| - |\mathcal{M}|$
$FN \leftarrow |G| - TP$
return TP, FP, FN

This design effectively prevents artificially inflated true positive counts caused by redundant predictions clustering around a single ground-truth GCI, which would otherwise lead to unrealistic improvements in performance metrics such as the F1 score. Under this matching strategy, the evaluation metrics are defined as follows:

- True Positive (TP): A predicted GCI that correctly matches a ground-truth GCI and is not used in multiple matches.
- False Positive (FP): A predicted GCI that does not match any ground-truth point.
- False Negative (FN): A ground-truth GCI that is not matched by any prediction.
- True Negative (TN): Not considered, as the overwhelming number of background points is not informative for analysis.

Based on the above definitions, we compute the following performance metrics:

- Precision:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{3}$$

- Recall:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{4}$$

- F1 score:

$$\text{F}_1 = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}. \tag{5}$$

The specific process of the GCI detection strategy under the given evaluation tolerance can be referred to **Algorithm 1**.

C. Label Construction and Enhancement

To enable data-driven models to learn GCIs more effectively, we design an enhanced labeling strategy that addresses the sparsity of GCI events and incorporates region-aware modifications for robust training across diverse corpora. The original GCI annotations are binary and sparse, with a value of 1 at each GCI location and 0 elsewhere. To mitigate the limitations imposed by such discrete and localized labels, this paper applies Gaussian smoothing along the temporal axis to generate soft labels. These labels retain a peak response at the GCI locations while introducing a smoothly decaying distribution around them. This formulation enables the model to capture local temporal dependencies and improves gradient flow during training.

Building upon this, we incorporate region-specific weighting adjustments based on additional annotations available in the training portion of the PTDB-TUG corpus. In this dataset, certain intervals are explicitly marked as ambiguous or weak, indicating either temporal uncertainty or low-amplitude glottal signals. These regions are down-weighted in both the hard and soft labeling paths. In the soft-label setting, Gaussian response amplitudes are attenuated according to the annotated region type; in the hard-label setting, binary activations are similarly suppressed. It is important to note that these region-based adjustments are applied only to the training set and are not used during validation or testing, ensuring fair and unbiased evaluation.

Finally, to support auxiliary tasks and improve the model's perception of phonation boundaries, we generate a voiced-region mask derived from the original GCI labels. This is done via morphological dilation and erosion operations, corresponding to a 50 *ms* window, producing a binary mask that broadly covers voiced segments. This auxiliary label facilitates the model's awareness of global glottal activity beyond individual closure instants.

Collectively, our labeling strategy enhances temporal continuity, integrates region-aware reliability cues, and introduces structured supervision for auxiliary tasks. These improvements substantially increase the robustness and generalization capability of the model when trained across corpora with varying levels of acoustic noise and annotation quality.

IV. The Proposed EGGNet Architecture

In this paper, we proposed a novel wide-context multi-branch convolutional model named EGGNet, which is designed for robust detection of GCIs from the EGG signals, as illustrated in Fig. 7. This model is optimized for generalization across datasets with varying acoustic conditions and label ambiguity, and it incorporates flexible signal preprocessing, multi-scale feature extraction, refinement, and joint supervision from hard and soft boundaries as well as auxiliary voiced-region masks.

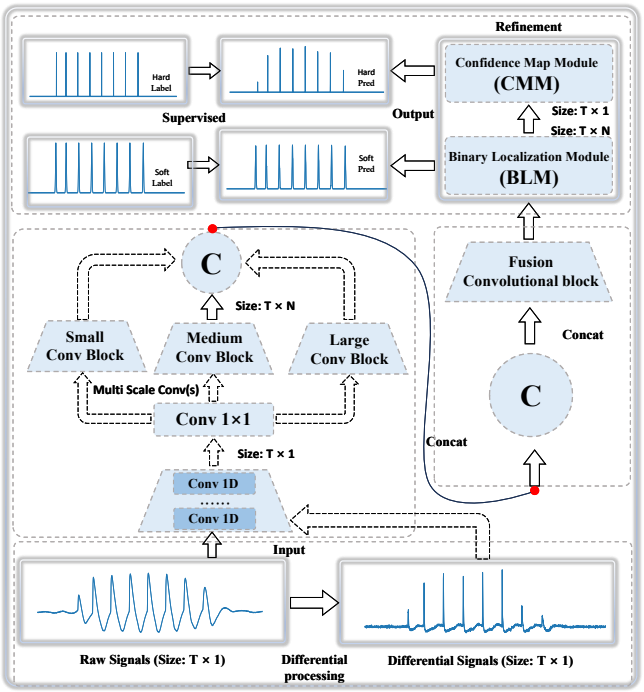

Fig. 7: Illustration of the EGGNet input filtering process, where T represents the number of signal sampling points and N represents the feature dimension.

A. Input Preprocessing

The proposed EGGNet accepts raw EGG signals as input and supports optional preprocessing steps, including band-pass filtering and first-order differential processing. For the input of the proposed EGGNet, two input channels are generated. The first input channel is the raw signal and the other input channel is the differentiated signal. Depending on the configuration parameter `input` set to `raw`, `diff`, or `hybrid`, the proposed EGGNet dynamically selects which input path and whether to apply filtering preprocessing. In particular, when the input configuration is `diff` or `hybrid`, the signal is then differentiated. Accordingly, the `raw` mode uses only the undifferentiated signal; the `diff` mode uses only the differentiated signal; the `hybrid` mode uses both. Moreover, all inputs are normalized before being processed by the network.

B. Multi-Scale Feature Encoding

For the feature encoding, the proposed EGGNet first employs a front-end module with two large-kernel convolutional layers on each input branch to extract low-level temporal structures from the time series of raw EGG. The extracted features are then processed by three parallel multi-scale convolutional subnetworks, each covering small, medium, or large receptive fields. In particular, each subnetwork consists of three 1D convolutional layers without downsampling, preserving the temporal resolution. In order to maintain consistent parameter counts in subsequent ablation studies, the multi-path control configuration of the proposed EGGNet uses three convolutional subnetworks with identical receptive fields, while the single-path control uses only one subnetwork. In both cases, the number of channels is adjusted to match the parameter count of the multi-scale setup. The outputs from all active branches are concatenated along the channel dimension, forming a unified temporal feature representation.

C. Refinement

To simultaneously model the probabilistic distribution of boundary regions and the precise localization of boundary positions, this paper proposes a novel cascaded supervision structure comprising two sequential convolutional modules, each addressing distinct modeling tasks, as illustrated in Fig. 7. The first-stage soft-label convolutional module, referred to as the confidence map module (CMM), is capable of generating a continuous confidence distribution. It is supervised by boundary probability targets smoothed with a Gaussian to enhance sensitivity to boundary regions. The second-stage hard-label convolutional module, referred to as the binary localization module (BLM), is designed for generating discrete boundary predictions aligned with binary location labels to ensure precise localization. In this case, these two supervisory signals complement each other, effectively balancing boundary sensitivity and localization accuracy.

Under the three training configurations of using soft-label only, hard-label only, and joint soft-hard supervision, the structure of these models remains completely consistent. Input features consistently pass through both convolutional modules sequentially, forming a unified forward path. Notably, when training with either hard-label-only or soft-label-only supervision, the loss is computed solely based on the output of the second convolutional module against the respective hard or soft labels. The core idea of this design ensures that the supervision method is the only experimental variable, effectively isolating it in ablation studies and eliminating confounding influences from variations in model capacity, path depth, or computational processes.

This method is designed to extract GCI events from the confidence sequence output by the model, which predicts the probability of an event at each time point. Given the high sparsity of GCI events in speech and the typically low model response at non-event positions, a fixed threshold of $\tau = 0.05$ is used to balance recall and false positive rates. As shown in **Algorithm 2**, the proposed EGGNet first applies local maxima filtering to select candidate peaks, retaining only those that exceed the threshold, fall within voiced segments, and correspond to local maxima. The final output $g(t)$ is a binary sequence with the same length as the input, where $g(t) = 1$

indicates a detected GCI event, whereas $g(t) = 0$ indicates no event.

Algorithm 2 Peak Extraction from Confidence Sequence

Require: Confidence sequence $s(t)$
Ensure: Binary GCI prediction $g(t)$
Set threshold $\theta \leftarrow 0.05$
$m(t) \leftarrow \text{MaxPool1D}(s(t), \text{kernel} = 5)$
Initialize $g(t) \leftarrow 0$ for all t
for all t **do**
  if $s(t) > \theta \wedge s(t) = m(t)$ **then**
    $g(t) \leftarrow 1$
  end if
end for
return $g(t)$

V. Performance Evaluation and Analysis

In this section, we conduct extensive comparative performance evaluation and comprehensive ablation experiments for our proposed EGGNet framework, and an in-depth analysis of the obtained experimental results has also been provided.

A. Performance Comparison Across Different Datasets

As described in Table IV, we have conducted a detailed comparison of the proposed EGGNet and the most widely used *peakdet* on the datasets CSTR-FDA, KEELE, and MOCHA-TIMIT, evaluating their performance in terms of Precision, Recall, and F1 metrics. Furthermore, we conducted a comprehensive set of ablation experiments. To clearly inform the reader about the specific variables we examined in the ablation experiments, we define the following abbreviations:

- **Single-Scale Convolution Block (SSCB)**: Contains three 16-channel convolutional blocks with the same receptive field, each consisting of three layers. The dimensions of each block are **[745, 373, 187]**. The total number of learnable parameters, including all convolutional filter weights and biases within this block, is 100.2 M.
- **Single-Path Convolution Block (SPCB)**: Contains a single 48-channel convolutional block with three layers. The dimensions of the block are **[401, 201, 101]**. The total number of learnable parameters, including all convolutional filter weights and biases within this block, is 100.4 M.
- **Multi-Scale Convolution Block (MSCB)**: Contains three 16-channel convolutional blocks with different receptive fields. The small, medium, and large convolution blocks have dimensions of **[321, 161, 81]**, **[641, 321, 161]**, and **[1281, 641, 321]**, respectively, as illustrated in Fig.7. The total number of learnable parameters, including all convolutional filter weights and biases within this block, is 100.6 M.

Moreover, in the ablation experiment comparison, depending on the input, we can set up three types of inputs: (1) **Differential signal input (DI)**; (2) **Raw signal input (RI)**; and (3) **Hybrid signal input (HI)**, where HI refers to the concatenation of the features from RI and DI with a band-pass filter sampled at 16000 Hz. The baseline schemes for ablation experiments can be referred to as **Simple Baselines**, which employ hard-coded labels and enable filters, as shown in Table II:

TABLE II: Description of the Designed **Simple Baselines** in Ablation Experiments.

Simple Baselines	Description
(Simple, SSCB, DI)	The SSCB mode takes DI as input signals.
(Simple, SPCB, DI)	The SPCB mode takes DI as input signals.
(Simple, SSCB, RI)	The SSCB mode takes RI as input signals.
(Simple, SPCB, RI)	The SPCB mode takes RI as input signals.
(Simple, SSCB, HI)	The SSCB mode takes HI as input signals.
(Simple, SPCB, HI)	The SPCB mode takes HI as input signals.

In addition to **Simple Baselines**, we have conducted comprehensive ablation experiments on EGGNet itself. For the optimal configuration, we refer to it as **(EGGNet, Optimal)**, which utilizes the RI type of input signal. Unlike *peakdet*, **(EGGNet, Optimal)** does not require differential processing and can accept raw signals. Furthermore, **(EGGNet, Optimal)** has introduced a hybrid label approach combining **Soft** and **Hard** labels for the first time, and in the **MSCB** mode, it has enabled the filter. Building upon **(EGGNet, Optimal)**, we propose the following baseline schemes, as shown in Table III.

TABLE III: Description of the Designed Baseline Schemes of **(EGGNet, Optimal)** in Ablation Experiments.

(EGGNet, Optimal)	Description
(EGGNet, DI)	The input signal is switched to DI.
(EGGNet, HI)	The input signal is switched to HI.
(EGGNet, Hard)	Using Hard Label
(EGGNet, Soft)	Using Soft Label
(EGGNet, w/o Filter)	Disable the filter.
(EGGNet, SPCB)	The SPCB mode is adopted.
(EGGNet, SSCB)	The SSCB mode is adopted.

As illustrated in Table IV, it can be easily observed that the state-of-the-art CGI detection method *peakdet* can achieve 99.7% Precision performance on the dataset CSTR-FDA, much lower Recall performance on the dataset KEELE, resulting in an unsatisfactory F1 score. In addition, the performance achieved by *peakdet* on the dataset MOCHA-TIMIT is extremely poor. Specifically, the Precision performance is only 48.9%, and the Recall performance is only 84.3%, leading to an F1 score of just 61.9%.

The primary reason for the above results is that the MOCHA-TIMIT dataset has a relatively lower SNR. For corpus datasets with a lower SNR, *peakdet* exhibits significant shortcomings. It erroneously identifies a large amount of white noise signals from silent moments as GCI, as illustrated in Fig. 2 of this paper. These results highlight that the most widely used *peakdet* exhibits its limited robustness to variations in signal quality.

For our ablation experiments of the **Simple Baselines** series, it can be seen that they achieve superior Precision, Recall, and F1 score performance on the CSTR-FDA and KEELE datasets with high SNR, particularly with the F1 score significantly outperforming *peakdet*. Additionally, it is noteworthy that they maintain relatively stable performance on the more noisy MOCHA-TIMIT dataset, especially with the Precision and F1 score performances far surpassing those of *peakdet*

To effectively overcome the challenges posed by the corpora with low SNR, we incorporate the MSCB mode and hybrid label approach combining Soft and Hard labels. These designs significantly enhance performance on the MOCHA-TIMIT dataset. As shown in the results of **(EGGNet, Optimal)**, we can observe that it achieves F1 scores of 96.3%, 96.2%, and 90.3% on the datasets CSTR-FDA, KEELE, and MOCHA-TRIMIT, respectively, significantly outperforming *peakdet*. Notably, **(EGGNet, Optimal)** achieves a performance gain of 45.88% in F1 score on the MOCHA-TIMIT dataset, which is characterized by high noise and low SNR. This represents a substantial performance improvement, largely addressing the technical challenges of GCI annotation for high-noise and low SNR corpus datasets.

Moreover, as demonstrated by the comparison results among **(EGGNet, Optimal)**, (EGGNet, DI), and (EGGNet, HI), interestingly, using the input type of RI outperforms the input types of DI and HI, inspired by *peakdet*, which suggests that the network of the proposed **(EGGNet, Optimal)** can effectively learn features directly from the waveform without the need for handcrafted representations. Moreover, it also shows that mixing DI and RI into HI is unnecessary, simplifying the model and reducing computational overhead.

From the results of (EGGNet, w/o Filter), it can also be observed that (EGGNet, w/o Filter) provides a slight advantage on the cleaner CSTR-FDA and KEELE datasets but performs poorly on the dataset of MOCHA-TIMIT, emphasizing the importance of filtering in noisy conditions. According to the results of (EGGNet, Soft), it can be seen that the soft-label variant achieves the highest Recall performance on the datasets of CSTR-FDA and MOCHA-TIMIT, which demonstrates that labeling smoothing enhances sensitivity to boundary-ambiguous samples. Based on this, we propose a cascade-label strategy in **(EGGNet, Optimal)**, wherein soft labels are used for coarse GCI localization, followed by hard labels for precise refinement. This balance between recall and precision is confirmed through ablation studies. In conclusion, multi-scale modeling (i.e., **MSCB**), label mechanism design (i.e., **Soft-Hard Combined Label**), and appropriate preprocessing collectively contribute to the stable and robust performance of the **(EGGNet, Optimal)** across a range of conditions.

B. Multi-scale feature analysis.

As shown in Fig. 8, we present the activation maps of the three receptive-field branches on the same speech segment. For the large-scale branch (i.e., `Feature: branch L`), it can be easily observed that the proposed EGGNet framework remains active even in silent intervals, which reveals the powerful capacity of the proposed EGGNet in sensing long-range segmental states and distinguishing voice-active from silent regions. For the medium-scale branch (i.e., `Feature: branch M`), it can be seen that the proposed EGGNet framework displays clear stripe-like activations aligned with the periodic vibration of the vocal folds, which indicates the strong modelling of the GCI rhythm. Furthermore, for the small-scale branch (i.e., `Feature: branch S`), the results illustrate that the proposed EGGNet framework yields the sparsest responses with sharp peaks tightly centred on individual GCIs, which highlights its sensitivity to fine-grained waveform details. As a result, the integration of these three scales endows the proposed EGGNet with a structurally organized division of labor. In particular, the large scale captures paragraph-level contextual information, the medium scale encodes periodic patterns, and the small scale precisely locates fine-grained temporal events. Together, they enhance the robustness and accuracy of GCI detection under diverse acoustic conditions.

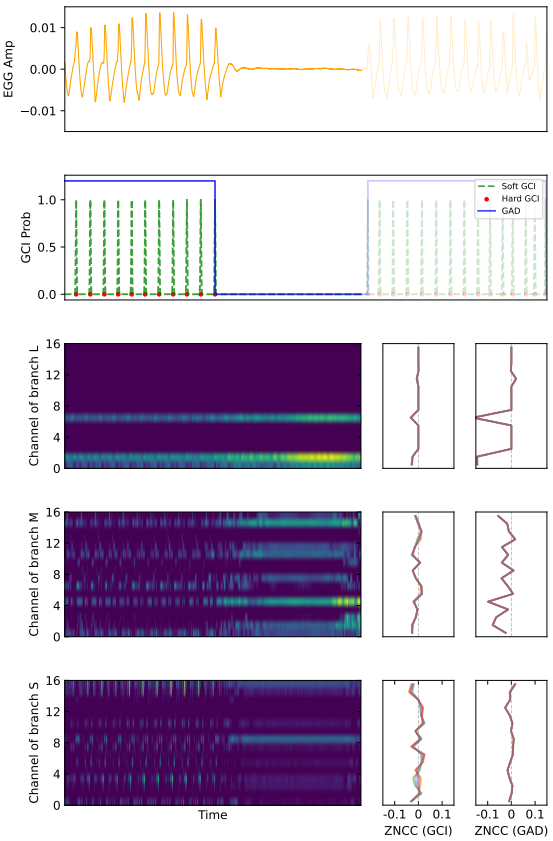

Fig. 8: The large branch captures silence context, the mid branch reflects periodicity, and the small branch shows sharp activations near GCIs for precise localisation.

Based on the preceding ablation results for SSCB and SPCB, it is evident that the multi-scale structure significantly outperforms single-path and single-receptive-field designs. Building on this, we further investigate the underlying mechanism by which multi-scale convolution enhances feature extraction. After obtaining the raw GCI labels, we apply a dilation-erosion operation to generate annotations for the vocal fold vibration intervals. This process serves a similar purpose to voice activity detection (VAD) and is referred to here as glottal activation detection (GAD), as indicated in Fig. 8. We then perform a zero-normalized cross-correlation (ZNCC) analysis between each channel output of the three convolution branches and the target GCI, as well as the corresponding GAD labels, over the lag set $\{-4, -2, 0, 2, 4\}$ [41]samples. In particular, the second column of the figure presents the ZNCCs with respect to the GCI, while the third column shows the NCCs with respect to the GAD, where multiple curves correspond to different lag values.

These results reveal that the output features of the large

TABLE IV: Performance comparison of model configurations across three datasets: CSTR-FDA, KEELE, and MOCHA-TIMIT. Values are mean(std) in %. Best and second-best metrics in each dataset are highlighted in red and yellow, respectively.

Model	CSTR-FDA (%)			KEELE (%)			MOCHA-TIMIT (%)		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
peakdet	99.7(–)	92.7(–)	96.0(–)	99.6(–)	79.5(–)	88.4(–)	48.9(–)	84.3(–)	61.9(–)
Simple Baselines									
(Simple, SSCB, DI)	97.6(3.1)	92.0(3.0)	94.7(1.5)	97.8(2.2)	91.3(3.6)	94.4(1.6)	83.3(8.0)	75.2(12.4)	78.1(6.0)
(Simple, SPCB, DI)	97.4(3.7)	93.4(1.1)	95.3(1.7)	98.1(2.1)	92.7(1.9)	95.3(1.2)	86.7(5.2)	77.5(4.5)	81.7(2.2)
(Simple, SSCB, RI)	98.3(1.0)	93.7(0.4)	96.0(0.6)	98.6(0.6)	93.1(0.5)	95.8(0.3)	84.2(4.5)	87.1(1.7)	85.6(2.0)
(Simple, SPCB, RI)	99.0(0.6)	94.3(0.7)	96.6(0.4)	98.1(1.7)	93.9(1.4)	95.9(1.1)	85.4(6.1)	84.3(3.4)	84.7(2.3)
(Simple, SSCB, HI)	94.3(6.0)	92.0(2.1)	93.0(3.5)	96.4(4.2)	91.2(1.6)	93.7(2.7)	81.6(4.5)	77.0(8.2)	78.9(4.1)
(Simple, SPCB, HI)	97.3(3.0)	91.9(4.6)	94.5(2.6)	98.1(1.6)	90.4(5.5)	94.0(3.1)	86.0(3.1)	77.0(4.2)	81.3(3.5)
(EGGNet, Optimal)	99.6(0.2)	93.8(0.6)	96.6(0.3)	99.3(0.2)	93.3(0.9)	96.2(0.4)	89.4(2.6)	91.1(1.4)	90.3(1.1)
(EGGNet, DI)	84.9(31.6)	90.9(6.7)	84.1(22.9)	83.7(33.4)	91.4(4.5)	83.2(25.3)	78.4(20.7)	86.8(6.5)	80.6(13.3)
(EGGNet, HI)	98.1(2.1)	93.2(0.8)	95.6(1.2)	98.2(2.0)	92.4(1.3)	95.2(1.5)	80.9(4.9)	86.1(5.1)	83.4(4.6)
(EGGNet, Hard)	98.1(2.2)	93.7(1.3)	95.8(1.3)	98.5(1.3)	92.3(2.9)	95.3(1.6)	86.0(6.9)	84.7(8.3)	84.9(3.3)
(EGGNet, Soft)	93.3(2.3)	95.6(0.4)	94.4(1.1)	93.5(1.8)	95.9(0.4)	94.7(0.8)	69.8(6.7)	95.5(1.0)	80.5(4.5)
(EGGNet, w/o Filter)	88.8(22.5)	84.2(24.5)	86.4(23.6)	94.2(8.0)	93.4(6.5)	93.8(7.2)	80.3(4.1)	87.3(19.5)	82.3(9.5)
(EGGNet, SPCB)	99.6(0.3)	93.4(0.9)	96.4(0.4)	99.2(0.3)	93.7(1.5)	96.4(0.7)	89.1(3.7)	88.0(3.1)	88.5(1.4)
(EGGNet, SSCB)	81.6(24.9)	91.6(1.2)	84.7(15.3)	79.4(27.7)	91.4(1.2)	82.6(16.9)	76.4(25.0)	81.5(3.1)	76.7(13.8)

branch are highly correlated with GAD, indicating that its long-range receptive field enables the model to determine whether the current moment lies within a vocal fold vibration interval. In contrast, the small branch outputs are primarily aligned with the GCI and exhibit slightly diverging line clusters over the lag set. For example, the noticeable variations in the amplitude and position of correlation peaks across different lags suggest greater sensitivity to small alignment offsets and thus better capability in capturing the instantaneous characteristics of GCIs. The middle branch strikes a balance between the two, achieving a trade-off between temporal coverage and instantaneous resolution.

C. T-SNE Analysis of Model Output Features and Soft Labels

In order to more intuitively explore the distribution of EGGNet's output features and soft labels, this paper employs t-distributed stochastic neighbor embedding (t-SNE) analysis [42], [43]. Physically, t-SNE aims to map high-dimensional features to lower dimensions (e.g., typically 2D or 3D) by preserving the similarity of probability distributions, thereby facilitating the visualization of the local structure and inter-class separation of the data. As shown in Fig. 9, we utilized t-SNE to visualize the output features of the proposed EGGNet, where the color represents the probability of the soft labels used for GCI detection. It is clearly observable that the output features of the proposed EGGNet model form well-clustered groups in the low-dimensional space, with distinct grouping phenomena evident. These clusters correspond to different categories of GCI events, revealing that the proposed EGGNet has successfully learned to distinguish various GCI instances during training. The soft labels are represented by a color gradient, indicating significant differences in confidence within EGGNet's predictions. Specifically, the color scale reflects the probability values assigned to each GCI instance, with lighter colors indicating higher confidence and darker colors indicating lower confidence in the model's predictions. This clustering effect supports the hypothesis that EGGNet can learn meaningful representations of the data, with high-dimensional features effectively mapped into separable groups. The use of soft labels allows for a more nuanced evaluation of model performance, enabling the detection of uncertainty and the differentiation between confident predictions and uncertain GCI instances. Consequently, the t-SNE analysis of soft labels not only confirms that EGGNet can effectively capture the temporal and structural features of GCI events but also provides insights into EGGNet's confidence in its predictions, which may be beneficial for future model optimization efforts.

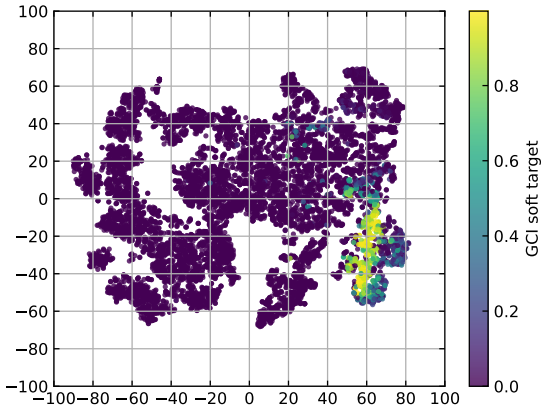

Fig. 9: T-SNE visualization of EGGNet's output features with soft label probability coloring for GCI detection, showing clear clustering and confidence variations.

D. Confusion Matrics of peakdet and EGGNet

As illustrated in Fig. 10, this paper uses confusion matrices to quantify the performance of the proposed EGGNet and *peakdet* on these three corpora, including CSTR-FDA, KEELE, and MOCHA-TIMIT, which provides a more intuitive understanding of the accuracy and robustness of EGGNet in GCI detection. As depicted in Fig. 10, the vertical axis of each subfigure represents the annotated F0 (i.e., Ground Truth in Hz), while the horizontal axis denotes the F0 calculated by EGGNet or *peakdet* based on the GCI sequence (i.e., Calculated in Hz). Both the vertical and horizontal axes cover a range of $30 \sim 270$ Hz. Ideally, the calculated results should be highly concentrated along the main diagonal of the confusion

matrix. The more concentrated the aggregation, the closer the model output is to the Ground Truth.

From Fig. 10 (a) and (d), it can be observed that on the CSTR-FDA dataset with high SNR, the confusion matrix shows that both *peakdet*and EGGNet are tightly clustered along the diagonal. This is consistent with the higher Precision, Recall, and F1 scores revealed in Table IV for the CSTR-FDA corpus, indicating that GCI extraction is relatively easy under ideal speech quality conditions. However, from Fig. 10 (b) and (e), on the KEELE dataset, compared to the proposed EGGNet, *peakdet* shows increased scatter outside the diagonal, with more errors occurring outside the diagonal. Combined with Table IV, it can be seen that the Recall performance significantly decreases at this point, while the confusion matrix of EGGNet remains tightly clustered along the diagonal, with only slight decreases in Precision, Recall, and F1, demonstrating its robustness under moderate noise conditions.

As illustrated in Fig. 10 (c) and (f), it can be seen that, on the MOCHA-TIMIT dataset with more complex and low SNR, *peakdet* exhibits a large number of errors outside the diagonal, with some non-GCI regions being misjudged as GCI, and an increase in missed detections. Combined with Table IV, it can be seen that the performance of *peakdet* in terms of Precision, Recall, and F1 score deteriorates significantly, with the F1 score being only 61.9%. In contrast, the confusion matrix of EGGNet remains dominated by the main diagonal, with Precision, Recall, and F1 values maintained at 89.4%, 91.1%, and 90.3%, respectively, significantly outperforming *peakdet*. As a result, across different corpora, the confusion matrix of EGGNet is highly consistent with the ideal diagonal form, and its performance is almost unaffected by changes in SNR. In contrast, the errors of *peakdet* significantly increase as signal quality decreases, especially on low SNR corpora. This analysis demonstrates that EGGNet not only accurately locates GCI but also maintains high robustness under high noise conditions. The MSCB modeling of EGGNet and the integration of soft and hard labels effectively enhance the accuracy and stability of detection.

VI. Conclusion

This work introduced EGGNet, an innovative framework named EGGNet for GCI detection from EGG signals, addressing limitations of conventional approaches such as *peakdet*. Through its multi-scale convolutional block design, cascaded soft–hard label supervision, and refined evaluation and matching strategies, EGGNet demonstrated superior robustness and accuracy across diverse corpora, achieving significant gains in low-SNR conditions. Comprehensive experiments and ablation studies validated its effectiveness, revealing consistent improvements in F1 scores and stable performance despite varying acoustic environments.

Looking forward, future research could extend EGGNet to more challenging real-world scenarios, including pathological speech assessment, multi-speaker interactions, and cross-language adaptability. Integrating EGGNet with end-to-end speech synthesis and enhancement systems could further improve the quality of voice source modeling. Moreover, exploring lightweight architectures and model compression techniques would facilitate deployment in embedded or mobile platforms, enabling real-time GCI detection in clinical, telecommunication, and human–computer interaction applications. These directions hold the potential to broaden the impact of EGGNet, making it a versatile and practical tool for both research and industry.

References

[1] M. Tanveer, A. Rastogi, V. Paliwal, M. Ganaie, A. Malik, J. Del Ser, and C.-T. Lin, "Ensemble deep learning in speech signal tasks: A review," *Neurocomputing*, vol. 550, p. 126436, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231223005593

[2] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *Information Fusion*, vol. 99, p. 101869, 2023.

[3] R. Feng, Y.-A. Chen, Y.-L. Liu, J.-H. Yuan, and Z.-H. 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.

[4] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Černockỳ, and D. Yu, "Neural target speech extraction: An overview," *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8–29, 2023.

[5] J. Yao, Q. Wang, Y. Lei, P. Guo, L. Xie, N. Wang, and J. 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.

[6] R. Feng, Y. Chen, Y. Hu, J. Du, and J. Yuan, "EGGCodec: A robust neural encodec framework for EGG reconstruction and F0 extraction," *arXiv preprint arXiv:2508.08924*, 2025.

[7] A. De Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[8] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 161–165.

[9] M. Mauch and S. Dixon, "PYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 659–663.

[10] V. Parsa and D. G. Jamieson, "A comparison of high precision Fo extraction algorithms for sustained vowels," *Journal of Speech, Language, and Hearing Research*, vol. 42, no. 1, pp. 112–126, 1999.

[11] W. Chu and A. Alwan, "Reducing F0 frame error of F0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3969–3972.

[12] J. Salamon, R. M. Bittner, J. Bonada, J. J. Bosch, E. Gómez, and J. P. Bello, "An analysis/synthesis framework for automatic F0 annotation of multitrack datasets." in *ISMIR*, 2017, pp. 71–78.

[13] P. C. Bagshaw, "Automatic prosodic analysis for computer aided pronunciation teaching," 1994. [Online]. Available: https://api.semanticscholar.org/CorpusID:60030269

[14] F. Plante, G. Meyer, and W. Ainsworth, "A pitch extraction reference database," *Children*, vol. 8, no. 12, pp. 30–50, 1995.

[15] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario." in *Interspeech*, 2011, pp. 1509–1512.

[16] S. S. Narayanan, E. Bresch, P. K. Ghosh, L. Goldstein, A. Katsamanis, Y. Kim, A. C. Lammert, M. I. Proctor, V. Ramanarayanan, Y. Zhu *et al.*, "A multimodal real-time mri articulatory corpus for speech research." in *Interspeech*, 2011, pp. 837–840.

[17] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 310–319, 2010.

[18] R. Islam, E. Abdel-Raheem, and M. Tarique, "Voice pathology detection using convolutional neural networks with electroglottographic (EGG) and speech signals," *Computer Methods and Programs in Biomedicine Update*, vol. 2, p. 100074, 2022.

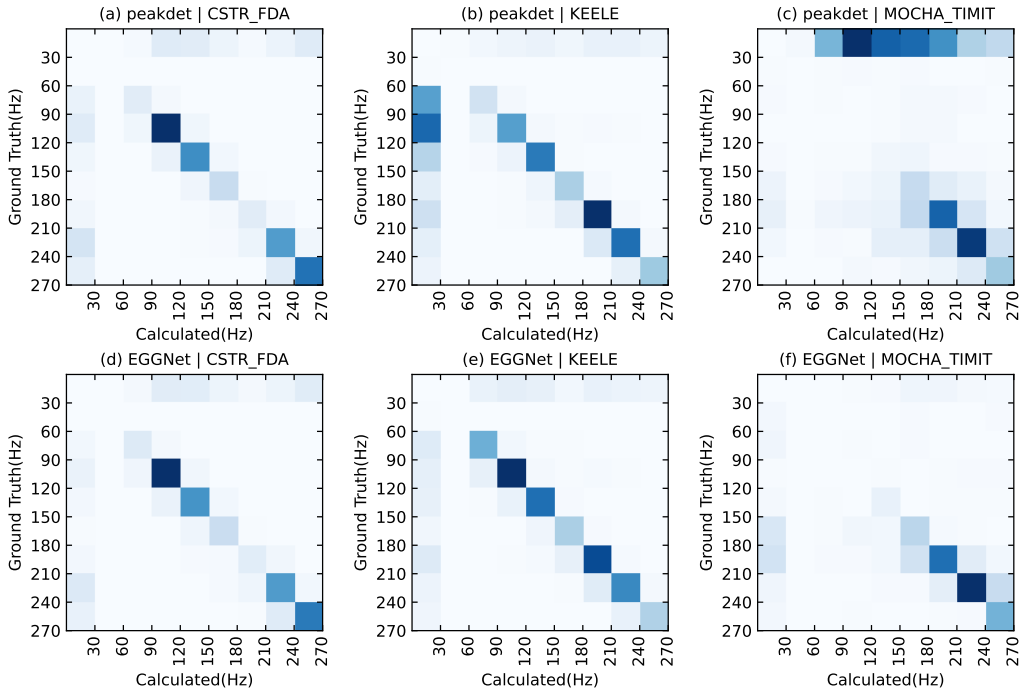

Fig. 10: The confusion matrix is used to quantify the performance of the proposed EGGNet and *peakdet* methods on these three corpora, CSTR-FDA, KEELE, and MOCHA-TIMIT.

[19] D. G. Childers and J. N. Larar, "Electroglottography for laryngeal function assessment and speech analysis," *IEEE Transactions on Biomedical Engineering*, vol. BME-31, no. 12, pp. 807–817, 1984.

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

[21] N. Henrich, C. d'Alessandro, B. Doval, and M. 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, 02 2004. [Online]. Available: https://doi.org/10.1121/1.1646401

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

[23] M. Mazaudon and A. 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.

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

[25] J. Z. Tomaszewska and A. Georgakis, "Electroglottography in medical diagnostics of vocal tract pathologies: A systematic review," *Journal of Voice*, 2023.

[26] J. Cabral, "Estimation of the asymmetry parameter of the glottal flow waveform using the electroglottographic signal." in *INTERSPEECH*, 2018, pp. 2997–3001.

[27] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP — a collaborative voice analysis repository for speech technologies," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 960–964.

[28] R. M. Bittner, B. McFee, J. Salamon, P. Li, and J. P. Bello, "Deep salience representations for F0 estimation in polyphonic music." in *ISMIR*. Suzhou (China), 2017, pp. 63–70.

[29] B. Bechtold, "Pitch of voiced speech in the short-time fourier transform: Algorithms, ground truths, and evaluation methods," Ph.D. dissertation, Universität Oldenburg, 2021.

[30] M.-C. Nguyen, "Exploring machine learning perspectives for electroglottographic signals," Ph.D. dissertation, LIG (Laboratoire informatique de Grenoble), 2023.

[31] R. Islam, E. Abdel-Raheem, and M. Tarique, "Deep learning based pathological voice detection algorithm using speech and electroglottographic (EGG) signals," in *2022 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, 2022, pp. 127–131.

[32] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *EUROSPEECH*, 1995. [Online]. Available: https://api.semanticscholar.org/CorpusID:15121762

[33] L. Sukhostat and Y. 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.

[34] M. K. Hasan, S. Hussain, M. H. Setu, and M. N. I. Nazrul, "Signal reshaping using dominant harmonic for pitch estimation of noisy speech," *Signal processing*, vol. 86, no. 5, pp. 1010–1018, 2006.

[35] K. Funaki, "On evaluation of the F0 estimation based on time-varying complex speech analysis." in *Interspeech*, 2010, pp. 637–640.

[36] D. Wang, C. Yu, and J. H. L. Hansen, "Robust harmonic features for classification-based pitch estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 952–964, 2017.

[37] C. Wang and S. Seneff, "Robust pitch tracking for prosodic modeling in telephone speech," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, vol. 3, 2000, pp. 1343–1346 vol.3.

[38] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 2494–2498.

[39] L. Shi, J. K. Nielsen, J. R. Jensen, M. A. Little, and M. G. Christensen, "Robust bayesian pitch tracking based on the harmonic model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1737–1751, 2019.

[40] K. Heryan, W. Reklewski, A. Szaflarski, M. Ordowski, P. Augustyniak, and M. Miśkowicz, "Sensitivity of QRS detection accuracy to detector temporal resolution," in *2021 Computing in Cardiology (CinC)*, vol. 48. IEEE, 2021, pp. 1–4.

[41] A. Kaso, "Computation of the normalized cross-correlation by fast fourier transform," *PloS one*, vol. 13, no. 9, p. e0203434, 2018.

[42] M. Wattenberg, F. Viégas, and I. Johnson, "How to use t-SNE effectively," *Distill*, vol. 1, no. 10, p. e2, 2016.

[43] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.