JASA ARTICLE

Leveraging laryngograph data for robust voicing detection in speech

Yixuan Zhang,[1,a)] Heming Wang,[1,b)] and DeLiang Wang[1,2,c)]
[1]Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, Ohio 43210, USA

ABSTRACT:
Accurately detecting voiced intervals in speech signals is a critical step in pitch tracking and has numerous applications. While conventional signal processing methods and deep learning algorithms have been proposed for this task, their need to fine-tune threshold parameters for different datasets and limited generalization restrict their utility in real-world applications. To address these challenges, this study proposes a supervised voicing detection model that leverages recorded laryngograph data. The model, adapted from a recently developed CrossNet architecture, is trained using reference voicing decisions derived from laryngograph datasets. Pretraining is also investigated to improve the generalization ability of the model. The proposed model produces robust voicing detection results, outperforming other strong baseline methods, and generalizes well to unseen datasets. The source code of the proposed model with pretraining is provided along with the list of used laryngograph datasets to facilitate further research in this area. © *2024 Acoustical Society of America*. https://doi.org/10.1121/10.0034445

(Received 25 June 2024; revised 10 October 2024; accepted 3 November 2024; published online 20 November 2024)
[Editor: Paavo Alku] Pages: 3502–3513

I. INTRODUCTION

A speech signal consists of voiced, unvoiced, and silent intervals or segments. Detecting whether speech is voiced is known as voicing detection. This task is a crucial step in pitch estimation and benefits various speech processing tasks, such as speaker recognition (Bai and Zhang, 2021), computational auditory scene analysis (Wang and Brown, 2006), and speech recognition (Atal and Rabiner, 1976; Zolnay *et al.*, 2002). In deep learning, precise voicing detection contributes to enriching training data with essential context and segmentation. This can be especially valuable in situations where annotated data are limited. With voicing information, the potential of deep learning in advancing speech processing tasks can be further leveraged. It is important to note that voicing detection is different from voice activity detection (VAD), which aims to determine the presence or absence of speech activity in an audio signal. In contrast, voicing detection concerns detecting the voiced portions of speech signals.

Voiced speech is produced by the vibration of the glottis, creating periodic or semi-periodic pulses of air that resonate through the vocal tract, while unvoiced speech is aperiodic and produced when air flows through a narrow constriction in a way to produce turbulence noise with no glottis vibration (Stevens, 1998). Therefore, periodicity is the determining factor for voiced and unvoiced segments. In English, voiced sounds include all vowels and voiced consonants such as /g/, /v/, and /z/, while unvoiced sounds include unvoiced consonants such as fricatives (e.g., /f/) and stops (e.g., /p/). Unvoiced speech accounts for approximately 20%–25% of all speech sounds in terms of both phoneme occurrence and segment duration (Hu and Wang, 2008), which highlights the significant role that unvoiced sounds play in speech utterances. Further description of voiced and unvoiced speech segments in English will be provided in Sec. II.

Various approaches have been proposed to address voicing detection, by analyzing the waveform or energy of the signal or by examining spectral characteristics such as the presence of harmonics and formants. Conventional methods (Amado and Vieira Filho, 2008; Bachu *et al.*, 2010; De Cheveigné and Kawahara, 2002; Haggard *et al.*, 1970; Hosoda *et al.*, 2023; Zahorian and Hu, 2008; Talkin and Kleijn, 1995; Van Immerseel and Martens, 1992; Wang *et al.*, 2022) include analyzing the short-term autocorrelation sequence (De Cheveigné and Kawahara, 2002), zero-crossing rate (Amado and Vieira Filho, 2008), and energy of the speech signal (Bachu *et al.*, 2010), as well as a combination of these features with classification techniques like thresholding (Talkin and Kleijn, 1995) and rule-based approaches (Wang *et al.*, 2022). Deep learning based approaches for voicing detection often treat the task as part of pitch tracking and train a multi-task model (Han and Wang, 2014; Morrison *et al.*, 2023; Subramani *et al.*, 2024; Tran *et al.*, 2020; Zhang *et al.*, 2022). Pitch contours obtained from laryngograph data are commonly used as the ground-truth for evaluating the performance of these algorithms. However, existing methods for voicing detection have limitations in robustness and generalizability. Signal processing methods are sensitive to various types of noise,

a)Email: zhang.7388@osu.edu
b)Email: wang.11401@osu.edu
c)Email: dwang@cse.ohio-state.edu

https://doi.org/10.1121/10.0034445

including device noise, and typically require *ad hoc* tuning of a voicing decision threshold for each dataset, i.e., lacking consistency (Amado and Vieira Filho, 2008; Bachu *et al.*, 2010; Talkin and Kleijn, 1995). While deep neural network (DNN) methods (Ardaillon and Roebel, 2019; Kim *et al.*, 2018; Morrison *et al.*, 2023; Singh *et al.*, 2021) can perform voicing detection on clean speech by applying a threshold to the estimated periodicity, the sensitivity of this threshold still presents a challenge. The optimal threshold can vary, depending on the dataset used. Furthermore, the limited training data can lead to generalization issues, resulting in unreliable performance.

To address the above issues with existing methods, we leverage multiple datasets that provide laryngograph recordings to train a voicing detection model. Specifically, the voicing detection model is adapted from CrossNet (Kalkhorani and Wang, 2024) and is trained using ground-truth labels derived from the laryngograph data. Voicing labels obtained from laryngograph recordings are generally considered the gold standard for evaluating the accuracy of voicing detection models. By gathering existing datasets that contain laryngograph recordings, we obtain an adequate amount of data to train the proposed model.

A laryngograph, also known as an electroglottograph, is a medical device used for measuring the electrical activity of the larynx during speech production. It is a non-invasive device that is placed on the skin of the neck and detects changes in the electrical impedance of the vocal folds as they vibrate by emitting a high-frequency electrical signal into the neck. These changes in impedance are used to create a waveform that represents the movement of the vocal folds during pronunciation. Compared to microphone recordings, laryngograph recordings have several advantages for producing accurate voicing decisions. First, the laryngograph provides a direct measure of the vibration of the vocal folds, which is the source of the voiced speech signal. This is more accurate than methods that rely on indirect measures of voicing, such as the spectral or temporal characteristics of the speech signal. Second, the laryngograph is relatively unaffected by variations in the amplitude or frequency of the speech signal caused by acoustic noise or interference, which can be a problem for voicing detection methods based on microphone recordings. This makes the laryngograph a reliable tool for detecting voicing, especially in adverse acoustic environments. On the other hand, conducting laryngograph recordings is a cumbersome job. As a result, publicly accessible laryngograph data are limited, particularly from the perspective of large-scale DNN training.

Even though the laryngograph is regarded as the gold standard, it has certain limitations in capturing voicing details. In voiced speech, the vocal folds open and close in a quasi-periodic manner, transforming the glottal airflow into a series of flow pulses, which correspond to the excitation source of the speech signal. The electroglottographic (EGG) signal provided by laryngograph, however, primarily reflects the vocal fold contacting area. The exact moments of vocal fold closure and opening cannot be precisely determined from the EGG signal (Herbst, 2020; Herbst *et al.*, 2014), which can lead to potential timing discrepancies between the EGG signal and the speech signal. Additionally, in soft-onset voicing, vocal fold vibration and airflow modulation often start before the EGG signal. Alternative recording devices like throat microphones may help to overcome such limitations (Sahidullah *et al.*, 2018).

This paper presents a robust voicing detection model for clean speech that achieves state-of-the-art performance by leveraging multiple laryngograph datasets for training. We find that the model trained on accessible laryngograph datasets already yields good generalization. To further mitigate potential generalization issues, we conduct pretraining on the large-scale Librispeech dataset (Panayotov *et al.*, 2015), which leads to improved and more robust voicing detection performance. The contributions of our work can be summarized as follows:

- We investigate the distinct characteristics of voiced and unvoiced speech sounds and assess the feasibility of training on laryngograph datasets.
- We develop a supervised voicing detector that can accurately estimate voicing in clean speech and show robust performance across different corpora.
- Unlike previous methods, we leverage pretraining on the Librispeech dataset to address the challenge of insufficient laryngograph data.
- We conduct a comprehensive evaluation of the proposed method and comparison against other strong baselines.
- We release a pip-installable PYTHON library containing the trained model, which can be used to generate reliable ground-truth labels in cases where laryngograph data is not available, along with compiled laryngograph datasets.

This paper is structured as follows. In Sec. II, we provide a description of voiced and unvoiced English speech sounds and their characteristics. Sections III and IV describe related works and publicly accessible laryngograph datasets. Section V presents our voicing detection model. Sections VI and VII describe our experimental setup and evaluations of the proposed approach, including comparisons with existing methods. Finally, Sec. VIII provides concluding remarks. The source code and pretrained model used in this study are provided at https://github.com/YIXUANZ/rvd.

II. VOICED AND UNVOICED SPEECH SOUNDS

How to distinguish between voiced and unvoiced speech sounds? As discussed in Sec. I, the primary characteristic is periodicity, which is evident as harmonic patterns in the frequency domain. Therefore, detecting frames with harmonic patterns in the spectrum of the speech signal becomes an intuitive approach. Nonetheless, this task can be difficult in certain scenarios. Unvoiced frames exhibit no harmonic patterns in their spectrum and can be challenging to distinguish from background noise. Although the presence of harmonic structure is a reliable indicator of a voiced frame, it can be still difficult to recognize such harmonic

patterns in frames located between voiced and unvoiced intervals due to co-articulation effects. In such cases, harmonic patterns may be ambiguous, even though harmonic components still exist in the signal. In order to characterize these ambiguous frames, one can utilize contextual cues to make a determination. Linguistic features of the language being spoken provide helpful clues in distinguishing voiced and unvoiced speech segments. In English, phonemes are classified as either voiced or unvoiced (Ladefoged, 2001). Table I provides a catalog of the voiced and unvoiced phonemes in English, where all vowels, approximants, and nasals are voiced (Ladefoged, 2001). Certain consonants, including stops, fricatives, and affricates, have pairs of voiced or unvoiced sounds. It should be noted that the phoneme /h/ can be pronounced in either a voiced or unvoiced way.

The use of a laryngograph provides an effective way to distinguish between voiced and unvoiced frames empirically. As explained in Sec. I, laryngograph recordings provide a direct measurement of the vibrations from the source of the voiced signal, which is relatively unaffected by amplitude or frequency variations caused by environmental noise or interference.

Figure 1 shows a comparison between the magnitude spectrogram of a microphone recording and a laryngograph waveform. The audio recording is from the FDA dataset (Bagshaw *et al.*, 1993) and corresponds to the utterance "When forced to make a choice, Sarah chose ping-pong as her favorite game." We can observe that the laryngograph spectrogram provides a clear distinction between voiced and unvoiced intervals of the speech signal. For example, the word "choice" can be observed in the spectrograms between 1.122 and 1.537 s. This word is composed of both voiced and unvoiced sounds, but the unvoiced sounds are not captured in the laryngograph. The microphone recording, however, exhibits both kinds of sound, potentially complicating voicing detection.

III. RELATED WORKS

Numerous studies have been conducted for voicing detection given its importance for applications such as speech synthesis. Earlier studies primarily focus on developing signal processing algorithms (Drugman and Alwan, 2011; Koutrouvelis *et al.*, 2016; Kumar and Rao, 2016; McAulay and Quatieri, 1990; Narendra and Rao, 2015; Talkin and Kleijn, 1995; Upadhyay and Pachori, 2015). Among these algorithms, the robust algorithm for pitch tracking (RAPT) (Talkin and Kleijn, 1995) and the summation of residual harmonics (SRH) (Drugman and Alwan, 2011) algorithm are considered as the standard methods in clean and noisy speech, respectively (Koutrouvelis *et al.*, 2016). RAPT is a time-domain method that employs the normalized cross-correlation function (NCCF) (Atal, 1972) and dynamic programming for pitch tracking. In the post-processing stage, a voicing decision is made by applying dynamic programming to select the set of NCCF peaks in frames containing voiced speech signals, or to make no selection otherwise. SRH (Drugman and Alwan, 2011) leverages harmonic information in the residual signal to estimate pitch and make voicing decisions. It calculates SRH using the amplitude spectrum of the residual signal. During unvoiced intervals of speech, SRH values tend to be lower. Therefore, the algorithm applies a simple local threshold to SRH values to make voicing decisions, and a speech frame is classified as voiced if its SRH value is above the threshold and unvoiced otherwise.

In recent years, there has been a growing interest in exploring deep learning approaches for voicing detection, but primarily focusing on noisy or multi-talker scenarios. These approaches aim to address voicing detection and pitch estimation simultaneously. For example, studies in Han and Wang (2014) and Liu and Wang (2018) treat the two tasks as a multi-class classification problem, while others (Tran *et al.*, 2020) employ a multi-task learning approach to jointly perform the two tasks. In these methods, ground-truth labels are obtained by applying a pitch tracker to the microphone recordings of clean speech, which limits the

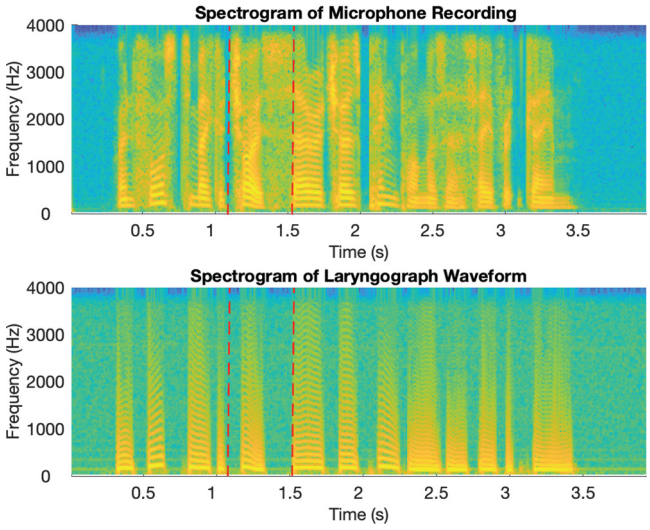

FIG. 1. (Color online) Magnitude spectrograms of microphone and laryngograph recordings of an utterance from a female speaker in the FDA dataset (Bagshaw *et al.*, 1993). The demarcated interval corresponds to the word "choice."

TABLE I. Voiced and unvoiced phonemes in English.

Phoneme type	Phoneme	Voiced or unvoiced?
Vowels	All	Voiced
Approximants	All	Voiced
Nasals	All	Voiced
Stops	/d/, /b/, /g/	Voiced
	/t/, /p/, /k/	Unvoiced
Fricatives	/z/, /v/, /ʒ/, /ð/	Voiced
	/s/, /f/, /ʃ/, /θ/	Unvoiced
	/h/	Both
Affricates	/dʒ/	Voiced
	/tʃ/	Unvoiced

https://doi.org/10.1121/10.0034445

accuracy of the trained model due to errors introduced by the pitch tracker on such clean speech. When it comes to ground truth, as discussed earlier, laryngograph data are considered to be the most reliable reference (Plante *et al.*, 1995). Several approaches have been proposed to address voicing detection in clean speech. Among them is the representative CREPE method (Kim *et al.*, 2018), which is trained on synthetic data to provide a periodicity estimate for each frame. The degree of periodicity indicates the likelihood of the presence of voiced speech within the frame, with higher values indicating a greater likelihood of voicing. Similar to SRH, a simple threshold is utilized to determine whether a frame is voiced. Such an approach sometimes produces unreliable voicing decisions. In Morrison *et al.* (2023), an entropy-based method is introduced for generating periodicity, and along with several training strategies, it significantly improves the accuracy of voicing decisions. Another approach involves utilizing a laryngograph to create annotations, which can then be employed to train a model on microphone recordings. For example, Drugman *et al.* (2018) incorporate both internal data and the CMU Arctic dataset (Kominek and Black, 2004) in their training data. Labels are obtained from laryngograph data, and a leave-one-speaker-out cross-validation scheme is employed during training to assess the effectiveness of their approach. While the idea is sensible, there is certainly room for improvement. First, their training dataset is relatively small, which potentially limits the generalizability of their trained model. Second, they employ a plain multi-layer perception (MLP), which may not be able to model complex patterns in the data as well as more advanced DNNs.

To our knowledge, there is currently no open-source DNN-based voicing detection algorithm trained on accessible laryngograph data, which hinders the effort of building on and improving earlier work. Our study intends to rectify this situation.

IV. LARYNGOGRAPH DATASETS, PREPROCESSING, AND LABEL GENERATION

A. Laryngograph datasets

Voicing labels generated from laryngograph recordings are widely used as ground-truths for evaluating voicing detection methods. Table II lists publicly accessible datasets employed in this study and provides relevant details for each dataset. We do not incorporate publicly accessible datasets that provide fewer than 100 utterances. Among the five datasets, three provide reference pitch and voicing labels extracted by different algorithms. FDA (Bagshaw *et al.*, 1993) is a relatively small dataset that provides microphone and laryngograph recordings from a male and a female speaker, and each speaker has 50 utterances. The provided reference labels in the FDA dataset are extracted using a "pulse" location algorithm where the duration between consecutive pulses are derived and converted to Hertz. If the value is within a certain range, the duration is considered voiced. Otherwise, it is considered unvoiced. PTDB-TUG (Pirker *et al.*, 2011) has ten male speakers and ten female speakers and around 4720 utterances in total. The provided reference labels are extracted by first applying a high-pass filter on laryngograph waveforms to remove low frequency components caused by larynx movements and then applying the RAPT (Talkin and Kleijn, 1995) algorithm on the filtered laryngograph waveforms. The KEELE (Plante *et al.*, 1995) dataset has recordings from five adult male speakers, five adult female speakers, and five children. We can only find the recordings from adult speakers. For male speakers, the length of each recording is from 27 s to 40 s. For female speakers, the length is from 28 s to 30 s. To better process the data, we further split the recordings to utterances around 3 s long. In total, we obtained 98 utterances. Note that the provided reference labels in the PTDB-TUG and KEELE datasets align with the 10 ms frame shift used in the voicing detection algorithms for evaluation (detailed in Sec. VII). The FDA dataset provides the start and end times of voiced intervals, and we generate labels with the 10 ms frame shift within these intervals. Mocha-TIMIT[1] and CMU Arctic (Kominek and Black, 2004) are relatively large datasets but do not provide reference labels. The Mocha-TIMIT dataset has 4028 utterances, which are uttered by four male and five female speakers. In the CMU Arctic dataset, we find that the recordings from two male speakers and one female speaker come with laryngograph waveforms. In total, the collected datasets contain 12 323 utterances from 22 male and 22 female speakers.

B. Data preprocessing

While the laryngograph data in these datasets are generally suitable for training purposes, our review reveals that two of the five datasets in Table II (PTDB-TUG and Mocha-TIMIT) contain problematic recordings. Upon examining

TABLE II. Description of accessible laryngograph datasets.

Dataset	Speaker information	No. of utterances	Label provided?	Label extraction method
FDA (Bagshaw *et al.*, 1993)	1 male and 1 female	100	Yes	Pulse location algorithm
PTDB-TUG (Pirker *et al.*, 2011)	10 males and 10 females	4720	Yes	RAPT Algorithm
KEELE (Plante *et al.*, 1995)	5 males and 5 females	98 (approximated)	Yes	Autocorrelation algorithm
Mocha-TIMIT[a]	4 males and 5 females	4028	No	
CMU Arctic (Kominek and Black, 2004)	2 males and 1 female	3377	No	
Total	22 males and 22 females	12 323		

[a]See https://data.cstr.ed.ac.uk/mocha.

the PTDB-TUG dataset, we find a number of laryngograph waveforms to be of low quality, such as the one depicted in Fig. 2(a). These recordings do not appear to capture vocal fold movements, making them unsuitable for extracting ground-truth labels. For the Mocha-TIMIT dataset, we have identified some files that contain noisy harmonic patterns in silence intervals, as illustrated in Fig. 2(b). These silence intervals would be recognized as voiced frames by a pitch extraction algorithm due to the presence of harmonic structure. To ensure evaluation accuracy, we perform manual exclusion and correction on these two datasets. On PTDB-TUG, the waveforms with spectrograms showing the patterns in Fig. 2(a) are excluded. On Mocha-TIMIT, we correct those waveforms that contain mildly noisy harmonic patterns where the RAPT algorithm can determine clear boundaries for voiced regions but incorrectly indicates voicing in silent intervals. Furthermore, we exclude the waveforms with strongly noisy harmonic patterns where even manual correction leads to errors; these scenarios typically involve voicing boundaries incorrectly detected by RAPT. As a result, a total of 1230 waveforms are excluded from the PTDB-TUG dataset, which originally comprises 4720 audio samples. Additionally, nearly 500 waveforms are corrected and around 270 waveforms are excluded for the Mocha-TIMIT dataset.

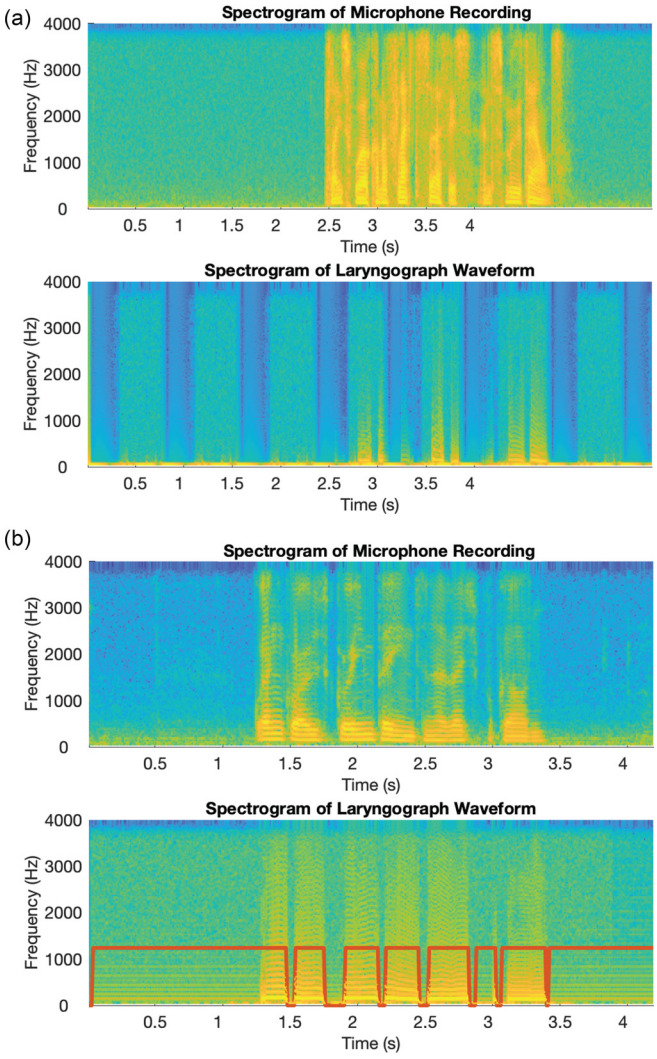

FIG. 2. (Color online) Examples of low-quality laryngograph data in (a) the PTDB-TUG dataset and (b) the Mocha-TIMIT dataset, with corresponding spectrograms of microphone and laryngograph recordings. The red line in panel (b) represents the reference voicing labels, including the erroneous labels extracted from the flawed laryngograph waveform.

C. Label generation

Given laryngograph data, different algorithms can be used to extract ground-truth voicing labels, and there is no standard way to perform label extraction. Although different algorithms produce similar results, the results differ to some extent. It is common that a paper announcing a laryngograph dataset provides reference labels and encourages users to generate their own reference labels (Pirker *et al.*, 2011; Plante *et al.*, 1995).

In alignment with the method outlined for reference voicing label generation in PTDB-TUG (Pirker *et al.*, 2011), we employ the following steps to derive reference voicing labels from laryngograph datasets:

- Preprocess each dataset and manually remove all utterances with quality issues.
- High-pass filter each utterance to remove the lower frequency components caused by larynx movements. Specifically, apply a linear phase Kaiser filter with parameters $\beta = 5$ and $n = 2400$ to laryngograph signals. For female speaker signals, the cutoff frequency is set to $f_c = 25$ Hz, and for the male speakers, $f_c = 15$ Hz.
- Apply the RAPT algorithm to filtered laryngograph signals to produce voicing decisions.
- If an audio frame is considered voiced, the reference label y_v is set to 1. Otherwise, it is set to 0.

How much do different label extraction algorithms differ? We use the above method to extract reference labels from PTDB-TUG, KEELE, and FDA datasets and compare them to the provided reference labels. Alignment is performed to maximize the match between the provided and extracted labels. The mismatch rates, which represent the percentage of mismatched frames to all frames are given in Table III. We observe that the mismatch rate is around 2% for all datasets. Figure 3 shows an example utterance with the top mismatch rate in the FDA dataset. The figure shows that the provided reference labels tend to under-label voiced intervals, and our method provides more balanced voicing decisions. More specifically, as illustrated in the figure, the intervals with relatively few harmonics tend to result in under-labeling in the provided reference labels. On the other hand, detecting short transitions between voiced phones tends to be challenging for our method. As pitch estimation is supposed to be conducted only in voiced intervals, different voicing labels will impact pitch estimation results. Over-labeling in particular is expected to degrade pitch estimation performance.

https://doi.org/10.1121/10.0034445

TABLE III. Mismatch rates between provided labels and self-generated labels.

Dataset	Mismatch rate (%)
FDA	1.89
KEELE	2.19
PTDB-TUG	1.90

V. MODEL DESCRIPTION

Our voicing detection model is adapted from CrossNet (Kalkhorani and Wang, 2024), a recently proposed speech separation model that achieves the state-of-the-art performance by capturing global, cross-band, and narrow-band correlations in the time-frequency domain. The network structure is illustrated in Fig. 4. Following Zhang *et al.* (2023), where better $F0$ estimation and voicing detection performance are observed in the complex domain rather than in the magnitude domain, and aligning with the setup in Kalkhorani and Wang (2024), we choose a complex-domain input feature comprising a concatenation of the real and imaginary parts of the complex short-time Fourier transform (STFT) of a speech signal,

$$X_{t,f} = \left[\Re(S_{t,f}), \Im(S_{t,f})\right], \tag{1}$$

where $S_{t,f}$ represents the STFT of the speech signal at time t and frequency f and $\Re$ and $\Im$ denote the real and imaginary parts, respectively. The speech signal is normalized by its variance before calculating its STFT.

As shown in Fig. 4, the model consists of an encoder layer, N CrossNet blocks, and a separate decoder layer. The encoder layer is a one-dimensional convolutional layer (Conv1D) that transforms the input feature from $2 \times F \times T$ to $C \times F \times T$, where C, F, and T are the number of hidden channels, the number of frequency bins, and the number of frames, respectively. Additionally, a random-chunk positional encoding (RCPE) method is employed to address the out-of-distribution problem commonly encountered in positional encoding approaches. RCPE shows improved generalization abilities for handling longer sequences. Specifically, RCPE is implemented by selecting a contiguous chunk of positional embedding vectors from a pre-computed positional encoding matrix during training. The pre-computed positional encoding matrix is defined as

$$\mathrm{PE}(t, 2i) = \sin\left(\frac{t}{10\,000^{2i/F\times H}}\right), \tag{2a}$$

$$\mathrm{PE}(t, 2i+1) = \cos\left(\frac{t}{10\,000^{2i/F\times H}}\right), \tag{2b}$$

and RCPE is selected by

$$\mathrm{RCPE}(T) = \begin{cases} \mathrm{PE}[\tau : \tau + T - 1, \ldots] & \text{if training,} \\ \mathrm{PE}[1 : T - 1, \ldots] & \text{else,} \end{cases} \tag{3}$$

where during training, a random index τ is drawn from $[1, T^{\max} - T + 1]$ and a chunk from index τ to $\tau + T - 1$ is selected. $T^{\max}$ is the maximum desired sequence length during inference. During validation or testing, the first T embedding vectors are selected. The RCPE feature is then added to the input features, which are subsequently used as input to the CrossNet blocks.

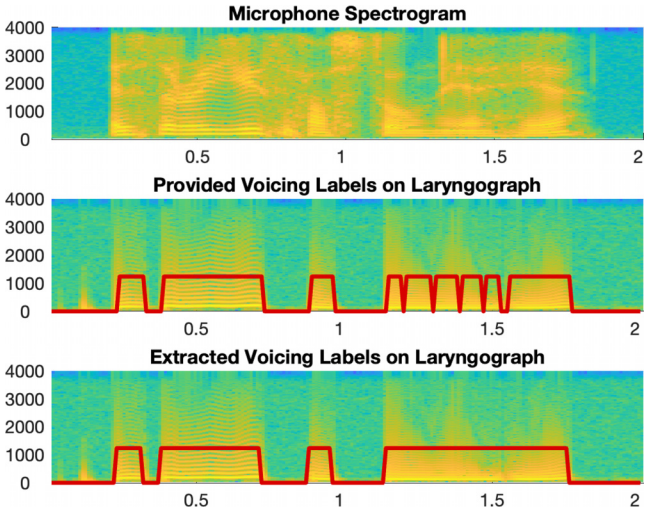

FIG. 3. (Color online) Comparison of provided and extracted voicing labels on the utterance with the highest mismatch rate in the FDA dataset (corresponding to utterance rl006 in the FDA dataset). The red line represents voicing decisions, where 0 indicates unvoiced and another positive value indicates voiced.

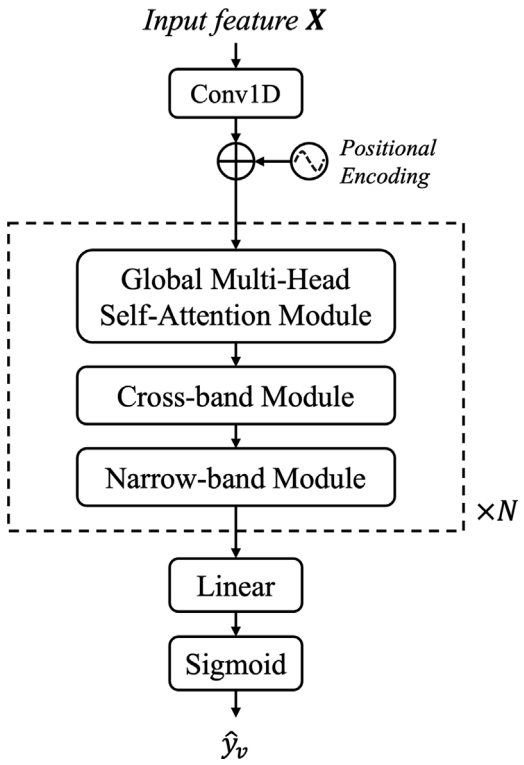

FIG. 4. Architecture of the voicing detection model based on CrossNet, with N representing the number of CrossNet blocks and $\hat{y}_v$ representing the voicing detection output ranging from 0 to 1.

Each CrossNet block consists of a global multi-head self-attention (GMHSA) module, a cross-band module, and a narrow-band module. The diagram illustrating each module within a CrossNet block is presented in Fig. 5. In the GMHSA module shown in Fig. 5(a), a point-wise two-dimensional convolutional (Point-wise Conv2D) layer is used to extract frame-level features. Its output with $L(2E + C/L)$ channels is split into L sets of queries Q, keys K, and values V, with dimensions of $E \times F \times T$, $E \times F \times T$, and $C/L \times F \times T$, respectively, and L represents the number of heads. Then, a self-attention layer is applied to learn global correlations. Outputs from all heads are then concatenated and sent to a Point-wise Conv2D layer with D output channels. This is followed by a parametric rectified linear unit (PReLU) activation function and layer normalization. Afterwards, the input to the GMHSA module is added to form the output of the module. The cross-band module, illustrated in Fig. 5(b), comprises two frequency-convolutional (Freq-Conv) modules and a full-band linear module. Each Freq-Conv module includes a layer normalization step, a grouped convolution layer along the frequency axis (F-GConv1D), and a PReLU activation function. In the full-band linear module, a linear layer is first used to reduce the number of hidden channels from C to C', followed by a sigmoid-weighted linear unit (SiLU) activation function. Next, several linear layers are applied along the frequency axis to extract full-band features, with their parameters shared by all the repeated CrossNet blocks to enhance parameter efficiency. Then, a linear layer followed by a SiLU activation function increases the number of hidden channels back to C. The input to the cross-band module is added to produce the output of the module. Figure 5(c) shows the narrow-band module, which is formed by a layer normalization, a linear layer followed by a SiLU activation, a time-convolutional (T-Conv) layer, and a final linear layer. The T-Conv layer consists of three grouped one-dimensional convolution (T-GConv1D) layers followed by a SiLU activation function, with the second T-GConv1D followed by a grouped normalization layer. The first linear layer in the narrow-band module maps the number of channels to C'', and the final layer maps the number of channels back to C. Finally, a linear layer followed by sigmoidal activation is employed to produce the probabilistic output $\hat{y}_v$.

In terms of network configuration details, we set the kernel sizes of the encoder layer, F-GConv1D, and T-GConv1D to 5, 3, and 5, respectively. We set the number of groups for F-GConv1D, T-GConv1D, and group normalization to 8. The model has $N = 7$ CrossNet blocks, with hidden channel sizes set to $C = 96$, $C' = 4$, and $C'' = 192$. The GMHSA module has four self-attention heads with an embedding dimension of $D = 64$ and $E = \lceil 512/F \rceil$, where $\lceil \cdot \rceil$ represents the ceiling operation.

To train the model and obtain the probabilistic output $\hat{y}_v$ for an estimated voicing decision, we minimize the binary cross-entropy loss $\mathcal{L}_v$ for voicing detection. The loss function is defined as

$$\mathcal{L}_v(y_v, \hat{y}_v) = -y_v \log \hat{y}_v - (1 - y_v) \log(1 - \hat{y}_v), \tag{4}$$

where y_v represents the binary ground-truth voicing label, with $y_v = 1$ denoting a voiced frame and 0 indicating otherwise.

To get voicing decisions during inference, $\hat{y}_v$ is compared against a threshold. The frame is decided as voiced if $\hat{y}_v$ is higher than the threshold and unvoiced otherwise. We consider two ways for determining the threshold. The first sets the threshold to 0.5 across all test sets, consistent with the probabilistic interpretation of voicing detection. In the second way, the threshold for each model is determined based on the receiver operating characteristic (ROC) curve, which plots the true positive rate (TPR) against the false positive rate (FPR) at different thresholds. We first determine the threshold that maximizes TPR – FPR for each corpus, then use the average of the optimal thresholds across all corpora.

VI. EXPERIMENTAL SETUP

A. Datasets

We train our model on five laryngograph datasets as described in Sec. IV: PTDB-TUG, Mocha-TIMIT, FDA, KEELE, and CMU Arctic. The PTDB-TUG and Mocha-TIMIT datasets were preprocessed using the method mentioned in Sec. IV B. Additionally, we utilize a dataset consisting of 50 000 utterances from the train-clean-360 subset of LibriSpeech (Panayotov *et al.*, 2015) for

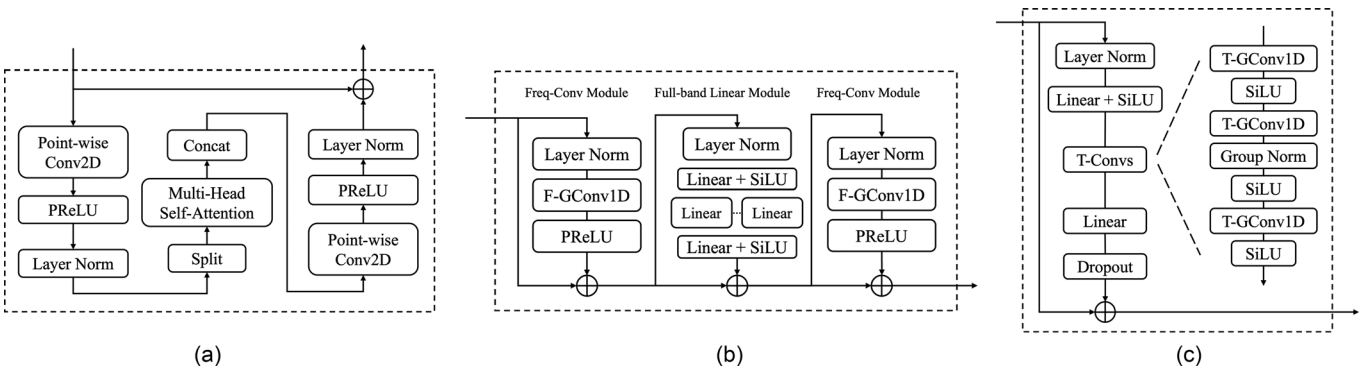

FIG. 5. Diagram of a CrossNet block. (a) Global multi-head self-attention module, (b) cross-band module, and (c) narrow-band module.

https://doi.org/10.1121/10.0034445

pretraining. The labels for these utterances are extracted using the RAPT algorithm. All audio files are downsampled to 8 kHz. For STFT computation, we use a Hamming window of 32 ms duration with a 10 ms frame shift.

B. Training methodology

To evaluate the performance of our approach, we employ a leave-one-corpus-out technique. Specifically, we divide the data from four of the five datasets into a training set, which comprises 90% of the data, and a validation set comprising the other 10%. The remaining dataset is used for testing or evaluation, and we repeat this process four times. By using this technique, we obtain a comprehensive assessment of the effectiveness of our method across multiple datasets, while minimizing the potential for bias and overfitting.

To enhance the generalizability of our trained model across different speakers and datasets, we employ a pretraining strategy. Specifically, we start with the model that has been trained on 50 000 microphone recordings from the LibriSpeech dataset. We use RAPT to generate pseudo-voicing labels for this pretraining. By incorporating this pretraining on LibriSpeech utterances that are more than four times those of the combined laryngograph datasets, we aim to improve the overall performance of the model on unseen data and speakers.

All models are trained with the Adam optimizer, with a maximum learning rate of 0.001. We use the PyTorch ReduceLROnPlateau scheduler and set the patience to 10 epochs and the reduction factor to 0.9. Gradient clipping is applied with a maximum value of 5 to avoid gradient explosion. We set the maximum training epoch number to 100, and all models converge within this limit.

C. Evaluation metrics

We evaluate the performance of voicing detection using voicing decision error (VDE), which indicates the percentage of frames that are wrongly classified in terms of voicing,

$$\mathrm{VDE} = \frac{N_{p\to n} + N_{n\to p}}{N}, \tag{5}$$

where N represents the total number of frames, $N_{p\to n}$ is the number of the voiced frames that are misclassified as non-voiced, and $N_{n\to p}$ is the number of non-voiced frames that are misclassified as voiced. We also use the F-measure ($F1$ score) to evaluate the performance of voicing detection, which is better suited for scenarios where the voiced and unvoiced labels are imbalanced.

D. Baselines and other evaluation details

Our evaluation includes quantitative comparisons against several strong baselines, including both signal processing and deep learning methods. For signal processing methods, we choose RAPT (Talkin and Kleijn, 1995) and SRH-Variant (Wang *et al.*, 2022), which is an improved version of SRH (Drugman and Alwan, 2011). It has been shown (Koutrouvelis *et al.*, 2016) that RAPT performs very well for clean speech, while SRH shows strong voicing detection performance for noisy speech.

For a deep learning baseline, we select a recent DNN-based approach called PENN (Morrison *et al.*, 2023), which is extended from the DNN methods of CREPE (Kim *et al.*, 2018), FCN (Ardaillon and Roebel, 2019), and DeepF0 (Singh *et al.*, 2021) and estimates the periodicity of each speech frame to classify it as voiced or unvoiced. Different from prior methods, PENN proposes a novel entropy-based method for extracting per-frame signal periodicity, which significantly enhances the classification accuracy of voiced and unvoiced speech frames.

For implementation of baselines, we use the code provided in the Speech Signal Processing Toolkit (SPTK) PYTHON package (SPTK working group, 2022) for RAPT and the original code provided in Wang *et al.* (2022) for SRH-Variant. For PENN, we use the default pretrained model provided in Morrison *et al.* (2023), which corresponds to FCNF0++ pretrained on the MDB-stem-synth and PTDB-TUG datasets, with a selected unvoiced threshold of 0.25 [see Sec. VI in Morrison *et al.* (2023)]. To ensure fair comparisons, we re-align the results from each baseline method for the lowest VDE.

In addition to the aforementioned baseline methods, we train a voicing detection model using the same training strategy as the proposed method but replace CrossNet by the DC-CRN architecture from our previous study (Zhang *et al.*, 2023). Like with CrossNet, DC-CRN generates a probabilistic output and is trained by minimizing the binary cross-entropy loss in Eq. (4). These experiments are conducted to assess the impact of DNN architecture and pretraining.

Using the second way of determining the voicing threshold described in Sec. V, we identify the following optimal thresholds for the proposed models: 0.48 for DC-CRN, 0.42 for DC-CRN with pretraining, 0.4 for CrossNet, and 0.5 for CrossNet with pretraining. In addition to the common threshold of 0.5, these model-specific optimal thresholds are used in the evaluations presented in Sec. VII.

VII. EVALUATIONS AND COMPARISONS

A. Cross-corpus results

To evaluate the proposed and baseline methods, we report a leave-one-corpus-out voicing detection results to assess cross-corpus generalization. The VDE and F1 results of the proposed methods and the baselines are given in Table IV, where the proposed methods are evaluated using the common threshold of 0.5. Table V shows the evaluation results of the proposed methods using the optimal thresholds determined through ROC analysis. It is worth noting that the PENN model is trained in part on PTDB-TUG, so some of the utterances in its PTDB-TUG evaluation are seen during training, resulting in potentially inflated results in this evaluation.

TABLE IV. Cross-corpus evaluation results in terms of VDE and $F1$ score.[a] Results in boldface represent the best results obtained for each test set.

Parameter	VDE (%)/$F1$ score for:				
Training set	M, K, F, C	P, K, F, C	P, M, F, C	P, M, K, C	P, M, F, K
Test set	PTDB-TUG	Mocha-TIMIT	KEELE	FDA	CMU Arctic
RAPT	3.47%/0.9230	10.41%/0.8618	5.75%/0.9490	4.61%/0.9407	6.46%/0.9429
SRH-Variant	5.39%/0.8648	13.53%/0.8239	9.37%/0.9113	7.84%/0.9042	8.79%/0.9190
PENN	2.37%/0.9446	5.24%/0.9230	12.31%/0.8835	10.05%/0.8466	5.22%/0.9516
DC-CRN	2.03%/0.9536	3.75%/0.9422	4.98%/0.9574	3.77%/0.9497	3.74%/0.9649
DC-CRN PT	1.83%/0.9580	3.43%/0.9451	**4.06%/0.9655**	4.32%/0.9418	3.47%/0.9675
CrossNet	1.84%/0.9574	3.75%/0.9442	5.02%/0.9562	**3.14%/0.9593**	3.68%/0.9652
CrossNet PT	**1.77%/0.9593**	**3.06%/0.9510**	4.36%/0.9620	3.96%/0.9451	**3.35%/0.9689**

[a]The proposed methods are evaluated using the common threshold of 0.5. The datasets used for training and evaluation are CMU Arctic (C), FDA (F), KEELE (K), Mocha-TIMIT (M), and PTDB-TUG (P). PT means with pretraining.

CrossNet and CrossNet with pretraining denote the versions of the proposed model with or without pretraining on LibriSpeech. Table IV shows that Mocha-TIMIT is the most challenging dataset for the signal processing methods, with the VDE rates of 10.41% for RAPT and 13.53% for SRH-Variant. This may be attributed to the fact that the recorded speech signals in Mocha-TIMIT have significant device noise, making voicing detection more difficult. On the other hand, the deep learning methods show more tolerance to such noise. PENN cuts the VDE of RAPT by half, and CrossNet with pretraining cuts the VDE by two thirds. It is worth noting that, as discussed in Koutrouvelis *et al.* (2016), RAPT has been shown to yield strong performance on clean speech, outperforming SRH-Variant on all datasets.

PENN, which was trained on a combination of PTDB and a synthetic dataset, outperforms the signal processing methods, except for the two small datasets of KEELE and FDA. In addition to the very low VDE on PTDB-TUG, which is partly due to some common utterances in training and testing, on the Mocha-TIMIT dataset, PENN achieves a VDE of 5.24% and a $F1$ score of 0.9230, much better than 10.41% and 0.8618 by RAPT. On the CMU Arctic dataset, PENN achieves a VDE of 5.22% and a $F1$ score of 0.9516, better than 6.46% and 0.9429 by RAPT. On the other hand, PENN performs poorly on the KEELE and FDA datasets, even worse than SRH-Variant, indicating a lack of generalization to these small datasets.

The proposed CrossNet model produces the best results across all datasets. As shown in Table IV, on Mocha-TIMIT and CMU Arctic, where PENN performs well, the CrossNet model obtains VDEs of 3.75% and 3.68%, respectively, compared to PENN's 5.24% and 5.22%. The corresponding $F1$ scores for CrossNet are 0.9442 and 0.9652, better than PENN's 0.9230 and 0.9516, respectively. On the small datasets where PENN does not perform well, the CrossNet model achieves VDEs of 5.02% for KEELE and 3.14% for FDA, and $F1$ scores of 0.9562 for KEELE and 0.9593 for FDA. These results are significantly better than those from RAPT. These results suggest that the trained CrossNet model has better generalization by leveraging multiple datasets. Furthermore, the CrossNet model exhibits outstanding performance on the PTDB-TUG dataset, even surpassing that of the PENN model that is trained in part on this corpus. It is also worth noting that directly training a larger CrossNet designed for speech separation (Kalkhorani and Wang, 2024) actually gives worse performance compared to the smaller model optimized for voicing detection. This is likely due to the limited number of utterances in the training set, which is insufficient for the original CrossNet model to learn voicing patterns without overfitting.

For the CrossNet model pretrained on the LibriSpeech dataset, our evaluation results demonstrate consistent improvements across datasets, with the exception of the small FDA corpus. For instance, pretraining improves the VDE on Mocha-TIMIT by 18.4% and on KEELE by

TABLE V. Cross-corpus evaluation results in terms of VDE and $F1$ score.[a] Results in boldface represent the best results obtained for each test set.

Parameter	VDE (%)/$F1$ score for:				
Training set	M, K, F, C	P, K, F, C	P, M, F, C	P, M, K, C	P, M, F, K
Test set	PTDB-TUG	Mocha-TIMIT	KEELE	FDA	CMU Arctic
DC-CRN	2.07%/0.9531	3.89%/0.9407	4.66%/0.9591	3.79%/0.9500	3.76%/0.9649
DC-CRN PT	1.87%/0.9574	3.76%/0.9420	**3.78%/0.9671**	3.77%/0.9489	3.53%/0.9674
CrossNet	1.85%/0.9575	4.05%/0.9408	4.92%/0.9572	**3.24%/0.9585**	3.60%/0.9663
CrossNet PT	**1.77%/0.9593**	**3.06%/0.9510**	4.36%/0.9620	3.96%/0.9451	**3.35%/0.9689**

[a]Models are evaluated using the optimal thresholds from ROC analysis. The datasets used for training and evaluation are CMU Arctic (C), FDA (F), KEELE (K), Mocha-TIMIT (M), and PTDB-TUG (P).

https://doi.org/10.1121/10.0034445

13.15% relatively, with $F1$ scores improved to 0.9510 and 0.9620, respectively. These outcomes suggest that pretraining the CrossNet model on LibriSpeech help to boost voicing detection performance across datasets.

In our previous study (Zhang *et al.*, 2023), we developed a DC-CRN model (Tan *et al.*, 2021) for both fundamental frequency ($F0$) estimation and voicing detection in noisy speech. From Table IV, we find that the DC-CRN based voicing detection model also shows good results and improves with pretraining, demonstrating the effectiveness of pretraining regardless of network architecture. The CrossNet based model with pretraining brings further improvements, especially on the relatively larger corpora of PTDB-TUG, Mocha-TIMIT, and CMU Arctic. For example, when tested on the CMU Arctic dataset, the DC-CRN model achieves a VDE of 3.74% and a $F1$ score of 0.9649. With pretraining, the performance of DC-CRN gets improved to a VDE of 3.47% and a $F1$ score of 0.9675. CrossNet with pretraining further reduces the VDE to 3.35% and increases the $F1$ score to 0.9689. In addition, CrossNet has much fewer trainable parameters: 1.56×10^6 compared to DC-CRN's 4.18×10^6.

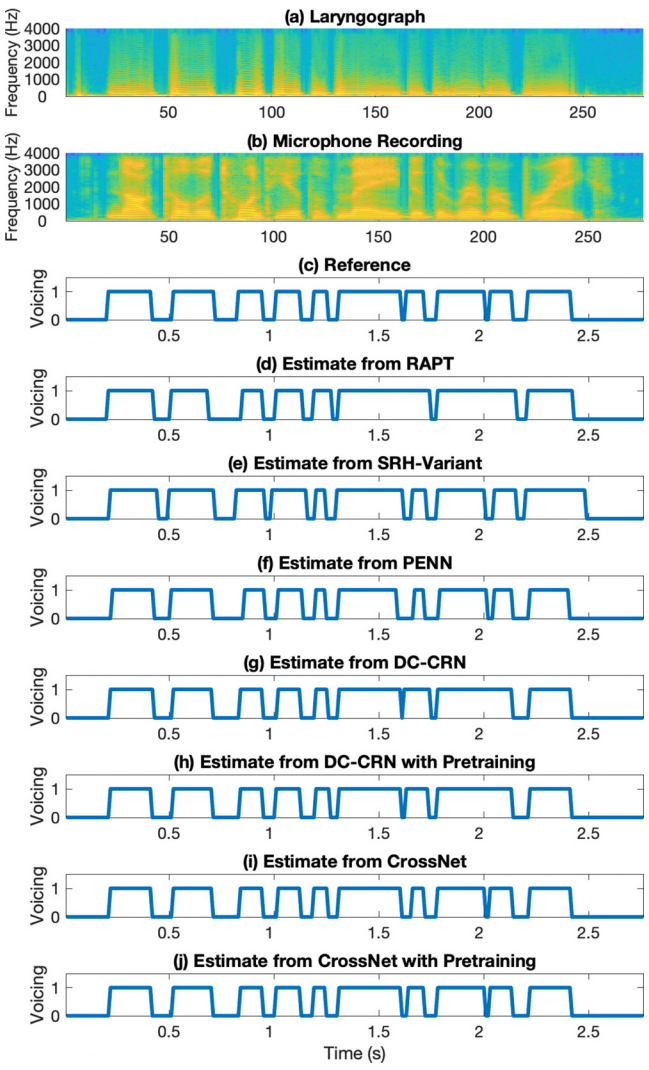

FIG. 6. (Color online) An example of voicing detection in clean speech, which is a male utterance ("Not till the twentieth of May did the river break.") from the CMU Arctic corpus. (a) Spectrogram of laryngograph waveform, (b) spectrogram of microphone recording, (c) reference voicing decisions, (d) estimated voicing decisions by RAPT, (e) estimated voicing decisions by SRH-Variant, (f) estimated voicing decisions by PENN, (g) estimated voicing decisions by DC-CRN, (h) estimated voicing decisions by DC-CRN with pretraining, (i) estimated voicing decisions by CrossNet, and (j) estimated voicing decisions by CrossNet with pretraining.

Table V presents the evaluation results of the proposed methods using the optimal thresholds determined through ROC analysis. By comparing these results with those in Table IV, we find that the identified optimal thresholds result in better performance on the smaller corpora of KEELE and FDA. In contrast, the common threshold shows slightly better performance on PTDB-TUG, Mocha-TIMIT, and CMU Arctic. Comparing the two tables shows that using the common threshold is simpler with comparable performance.

Figure 6 illustrates voicing detection performed on a laryngograph recording, specifically a male utterance from the CMU Arctic corpus. The proposed methods, including DC-CRN, DC-CRN with pretraining, CrossNet, and CrossNet with pretraining, are evaluated using the identified optimal thresholds. Figure 6(c) shows the reference voicing labels derived from Fig. 6(a) using the method described in Sec. IV C. As shown in Figs. 6(d) and 6(e), RAPT is prone to overestimating voiced regions, while SRH-Variant has both overestimation and underestimation errors. The DNN baseline, PENN, makes quality estimation but does not eliminate the underestimation problem. DC-CRN and DC-CRN with pretraining provide good estimates, although both slightly overestimate voiced frames around the 2-s point. The proposed CrossNet and CrossNet with pretraining models show better voicing detection performance than PENN and DC-CRN based models, yielding the most accurate estimates among all the methods.

B. Cross-corpus results on provided labels

As explained before, three of the five laryngograph corpora provide voicing labels. We now evaluate the proposed

TABLE VI. Cross-corpus evaluation results in terms of VDE and $F1$ score evaluated on datasets with provided labels.[a] Results in boldface represent the best results obtained for each test set.

Parameter	VDE (%)/$F1$ score for:		
Training set	M, K, F, C	P, M, F, C	P, M, K, C
Test set	PTDB-TUG	KEELE	FDA
RAPT	4.29%/0.9056	4.52%/0.9531	4.85%/0.9402
SRH-Variant	7.24%/0.8396	7.84%/0.9303	8.01%/0.8972
PENN	3.20%/0.9254	13.53%/0.8705	12.49%/0.8249
DC-CRN	2.97%/0.9326	4.66%/0.9539	4.73%/0.9390
DC-CRN PT	2.76%/0.9370	4.55%/0.9544	5.34%/0.9305
CrossNet	2.73%/0.9375	**4.26%/0.9577**	**4.34%/0.9452**
CrossNet PT	**2.69%/0.9387**	4.66%/0.9536	5.48%/0.9270

[a]The proposed methods are evaluated using the common threshold of 0.5. The datasets used for training and evaluation are CMU Arctic (C), FDA (F), KEELE (K), Mocha-TIMIT (M), and PTDB-TUG (P).

TABLE VII. Cross-corpus evaluation results in terms of VDE and $F1$ score evaluated on datasets with provided labels.[a] Results in boldface represent the best results obtained for each test set.

Parameter	VDE (%)/$F1$ score for:		
Training set	M, K, F, C	P, M, F, C	P, M, K, C
Test set	PTDB-TUG	KEELE	FDA
DC-CRN	2.93%/0.9336	4.59%/0.9547	4.95%/0.9363
DC-CRN PT	2.78%/0.9370	3.96%/0.9610	5.09%/0.9335
CrossNet	2.71%/0.9382	**3.95%/0.9612**	**4.28%/0.9467**
CrossNet PT	**2.69%/0.9387**	4.66%/0.9536	5.48%/0.9270

[a]Models are evaluated using the optimal thresholds from ROC analysis. The datasets used for training and evaluation are CMU Arctic (C), FDA (F), KEELE (K), Mocha-TIMIT (M), and PTDB-TUG (P).

and baseline methods using the provided labels, and the results are given in Table VI, where the proposed methods are evaluated using the common threshold of 0.5. It should be noted that, for the KEELE dataset, the provided labels include three classes: voiced, unvoiced, and uncertain. Our evaluation does not consider the uncertain frames for evaluation, which results in a lower VDE for KEELE compared to the results shown in Table IV. It is observed that the two signal processing methods have comparable VDE rates across provided and generated labels, indicating the consistency and similarity of the labels generated using laryngograph data. The mismatch in label generation methods between the training and test sets leads to a performance drop in the proposed methods when evaluated using the provided reference labels. In addition, the poor performance of PENN on the KEELE and FDA datasets confirms our earlier observation of its limited generalization ability. Improvements are observed in the DC-CRN and CrossNet models compared to the baselines, with pretraining showing benefits on the large PTDB-TUG dataset. The VDE of the CrossNet model is 1.56% lower than that of RAPT on PTDB-TUG, and the results of CrossNet and RAPT are similar on the smaller KEELE and FDA datasets. It is also observed that pretraining appears not to be beneficial in Table VI on the small KEELE and FDA datasets, due in part to the mismatch between the training and test data. These results demonstrate that the proposed CrossNet model and pretraining have strong generalizability across different label generation algorithms. We also evaluate the proposed models using the optimal thresholds determined through ROC analysis, with the results shown in Table VII. Overall, these results demonstrate slightly better performance compared to those in Table VI on the KEELE and FDA datasets.

VIII. CONCLUDING REMARKS

This study introduces a robust DNN-based voicing detection model for clean speech by using laryngograph data for training. The model employs a CrossNet architecture and incorporates a pretraining strategy on the LibriSpeech dataset. Our cross-corpus evaluations demonstrate that the proposed model outperforms signal processing and deep learning baseline methods and shows strong generalization. By open sourcing the model and the data, we expect to accelerate the progress of voicing detection and related research such as pitch tracking in challenging environments.

Given the success of using synthesized speech data for training DNN models for pitch estimation (see, e.g., Kim *et al.*, 2018), should voicing detection utilize synthetic speech data? While synthetic data allow complete control of ground-truth voicing labels, our preliminary investigation suggests that using synthetic data is not effective for voicing detection. A possible reason is that a voicing detection model trained on synthetic data generated from microphone recordings relies heavily on the similarity of the synthetic training data and the voicing patterns of microphone recordings, which is difficult to maintain at the boundaries between voiced and unvoiced intervals. Also, voicing detection from the microphone recording of a speech utterance is prone to errors, as highlighted in this paper. On the other hand, laryngograph data represent the gold standard for $F0$ and voicing label generation. Prior studies on pitch estimation (Ardaillon and Roebel, 2019; Zhang *et al.*, 2022) prefer synthetic data over laryngograph data partly because octave errors affect the accuracy of $F0$ labels derived from laryngograph recordings. This concern, however, does not extend to voicing labels.

In future work, we plan to apply the proposed voicing detection model to improve pitch tracking performance under adverse acoustic conditions, including background noise, room reverberation, and concurrent speakers.

ACKNOWLEDGMENTS

This work was supported in part by a National Institute on Deafness and Other Communication Disorders (NIDCD) grant (R01DC012048) and the Ohio Supercomputer Center.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are available within the article (Bagshaw *et al.*, 1993; Kominek and Black, 2004; Pirker *et al.*, 2011; Plante *et al.*, 1995) and at https://data.cstr.ed.ac.uk/mocha and https://github.com/YIXUANZ/rvd.

[1]Available at https://data.cstr.ed.ac.uk/mocha/.

Amado, R. G., and Vieira Filho, J. (**2008**). "Pitch detection algorithms based on zero-cross rate and autocorrelation function for musical notes," in *2008 International Conference on Audio, Language, and Image Processing (ICALIP)*, Shanghai, China (IEEE, New York), pp. 449–454.

Ardaillon, L., and Roebel, A. (**2019**). "Fully-convolutional network for pitch estimation of speech signals," in *Interspeech 2019*, Graz, Austria (ISCA, Stockholm, Sweden), pp. 2005–2009.

Atal, B., and Rabiner, L. (**1976**). "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," IEEE Trans. Acoust, Speech, Signal Process. **24**, 201–212.
Atal, B. S. (**1972**). "Automatic speaker recognition based on pitch contours," J. Acoust. Soc. Am. **52**, 1687–1697.
Bachu, R. G., Kopparthi, S., Adapa, B., and Barkana, B. D. (**2010**). "Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy," in *Advanced Techniques in Computing Sciences and Software Engineering*, edited by K. Elleithy (Springer, New York), pp. 279–282.
Bagshaw, P. C., Hiller, S. M., and Jack, M. A. (**1993**). "Enhanced pitch tracking and the processing of F_0 contours for computer aided intonation teaching," in *Eurospeech 1993*, Berlin, Germany (ISCA, Stockholm, Sweden), pp. 1003–1006.
Bai, Z., and Zhang, X. L. (**2021**). "Speaker recognition based on deep learning: An overview," Neural Netw. **140**, 65–99.
De Cheveigné, A., and Kawahara, H. (**2002**). "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am. **111**, 1917–1930.
Drugman, T., and Alwan, A. (**2011**). "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Interspeech 2011*, Florence, Italy (ISCA, Stockholm, Sweden).
Drugman, T., Huybrechts, G., Klimkov, V., and Moinet, A. (**2018**). "Traditional machine learning for pitch detection," IEEE Signal Process. Lett. **25**, 1745–1749.
Haggard, M., Ambler, S., and Callow, M. (**1970**). "Pitch as a voicing cue," J. Acoust. Soc. Am. **47**, 613–617.
Han, K., and Wang, D. L. (**2014**). "Neural network based pitch tracking in very noisy speech," IEEE/ACM Trans. Audio Speech Lang. Process. **22**, 2158–2168.
Herbst, C. T. (**2020**). "Electroglottography—An update," J. Voice **34**, 503–526.
Herbst, C. T., Lohscheller, J., Švec, J. G., Henrich, N., Weissengruber, G., and Fitch, W. T. (**2014**). "Glottal opening and closing events investigated by electroglottography and super-high-speed video recordings," J. Exp. Biol. **217**, 955–963.
Hosoda, Y., Kawamura, A., and Iiguni, Y. (**2023**). "Complex-domain pitch estimation algorithm for narrowband speech signals," IEEE/ACM Trans. Audio. Speech Lang. Process. **31**, 2067–2078.
Hu, G., and Wang, D. L. (**2008**). "Segregation of unvoiced speech from nonspeech interference," J. Acoust. Soc. Am. **124**, 1306–1319.
Kalkhorani, V. A., and Wang, D. L. (**2024**). "CrossNet: Leveraging global, cross-band, narrow-band, and positional encoding for single- and multi-channel speaker separation," arXiv:2403.03411 (Last viewed September 2024).
Kim, J. W., Salamon, J., Li, P., and Bello, J. P. (**2018**). "CREPE: A convolutional representation for pitch estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada (IEEE, New York), pp. 161–165.
Kominek, J., and Black, A. W. (**2004**). "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, Pittsburgh, PA (ISCA, Stockholm, Sweden), pp. 223–224.
Koutrouvelis, A. I., Kafentzis, G. P., Gaubitch, N. D., and Heusdens, R. (**2016**). "A fast method for high-resolution voiced/unvoiced detection and glottal closure/opening instant estimation of speech," IEEE/ACM Trans. Audio Speech Lang. Process. **24**, 316–328.
Kumar, S. S., and Rao, K. S. (**2016**). "Voice/non-voice detection using phase of zero frequency filtered speech signal," Speech Commun. **81**, 90–103.
Ladefoged, P. (**2001**). *Vowels and Consonants* (Blackwell, Oxford, UK), pp. 211–212.
Liu, Y., and Wang, D. L. (**2018**). "Permutation invariant training for speaker-independent multi-pitch tracking," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada (IEEE, New York), pp. 5594–5598.
McAulay, R. J., and Quatieri, T. F. (**1990**). "Pitch estimation and voicing detection based on a sinusoidal speech model," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Albuquerque, NM (IEEE, New York), pp. 249–252.
Morrison, M., Hsieh, C., Pruyne, N., and Pardo, B. (**2023**). "Cross-domain neural pitch and periodicity estimation," arXiv:2301.12258 (Last viewed September 2024).
Narendra, N., and Rao, K. S. (**2015**). "Robust voicing detection and F_0 estimation for HMM-based speech synthesis," Circuits Syst. Signal Process. **34**, 2597–2619.
Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (**2015**). "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia (IEEE, New York), pp. 5206–5210.
Pirker, G., Wohlmayr, M., Petrik, S., and Pernkopf, F. (**2011**). "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Interspeech 2011*, Florence, Italy (ISCA, Stockholm, Sweden), pp. 1509–1512.
Plante, F., Meyer, G., and Ainsworth, W. (**1995**). "A pitch extraction reference database," in *Eurospeech 1995*, Madrid, Spain (ISCA, Stockholm, Sweden), pp. 837–840.
Sahidullah, M., Thomsen, D. A. L., Hautamäki, R. G., Kinnunen, T., Tan, Z.-H., Parts, R., and Pitkänen, M. (**2018**). "Robust voice liveness detection and speaker verification using throat microphones," IEEE/ACM Trans. Audio Speech Lang. Process. **26**, 44–56.
Singh, S., Wang, R., and Qiu, Y. (**2021**). "DeepF0: End-to-end fundamental frequency estimation for music and speech signals," in *ICASSP 2021—International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada (IEEE, New York), pp. 61–65.
SPTK working group (**2022**). "Speech Signal Processing Toolkit (SPTK) (version 0.1.21) [computer program]," https://pysptk.readthedocs.io/en/latest/.
Stevens, K. N. (**1998**). *Acoustic Phonetics* (MIT Press, Cambridge, MA).
Subramani, K., Valin, J.-M., Büthe, J., Smaragdis, P., and Goodwin, M. (**2024**). "Noise-robust DSP-assisted neural pitch estimation with very low complexity," in *ICASSP 2024—International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Republic of Korea (IEEE, New York), pp. 11851–11855.
Talkin, D., and Kleijn, W. B. (**1995**). "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, edited by W. B. Kleijn and K. K. Paliwal (Elsevier Science, Amsterdam, Netherlands), Chap. 14, pp. 495–518.
Tan, K., Zhang, X., and Wang, D. L. (**2021**). "Deep learning based real-time speech enhancement for dual-microphone mobile phones," IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 1853–1863.
Tran, D. N., Batricevic, U., and Koishida, K. (**2020**). "Robust pitch regression with voiced/unvoiced classification in nonstationary noise environments," in *Interspeech 2020*, Shanghai, China (ISCA, Stockholm, Sweden), pp. 175–179.
Upadhyay, A., and Pachori, R. B. (**2015**). "Instantaneous voiced/non-voiced detection in speech signals based on variational mode decomposition," J. Franklin Inst. **352**, 2679–2707.
Van Immerseel, L. M., and Martens, J.-P. (**1992**). "Pitch and voiced/unvoiced determination with an auditory model," J. Acoust. Soc. Am. **91**, 3511–3526.
Wang, D., Wei, Y., Wang, Y., and Wang, J. (**2022**). "A robust and low computational cost pitch estimation method," Sensors **22**, 6026.
Wang, D. L., and Brown, G. J. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley-IEEE Press, Hoboken, NJ).
Zahorian, S. A., and Hu, H. (**2008**). "A spectral/temporal method for robust fundamental frequency tracking," J. Acoust. Soc. Am. **123**, 4559–4571.
Zhang, Y., Wang, H., and Wang, D. L. (**2022**). "Densely-connected convolutional recurrent network for fundamental frequency estimation in noisy speech," in *Interspeech 2022*, Incheon, South Korea (ISCA, Stockholm, Sweden), pp. 401–405.
Zhang, Y., Wang, H., and Wang, D. L. (**2023**). "*F*0 estimation and voicing detection with cascade architecture in noisy speech," IEEE/ACM Trans. Audio Speech Lang. Process. **31**, 3760–3770.
Zolnay, A., Schlüter, R., and Ney, H. (**2002**). "Robust speech recognition using a voiced-unvoiced feature," in *Proceedings of the 7th International Conference on Spoken Language Processing. ICSLP 2002*, Denver, CO (ISCA, Stockholm, Sweden), pp. 1065–1068.