CTC 강제 정렬을 활용한 음성 휴지 구간 탐색

구다연, 박화영, 김홍국*

*광주과학기술원, *(주)오니온에이아이

{dayeonku, hwayoung_park}@gm.gist.ac.kr, *hongkook@gist.ac.kr

Speech Pause Detection Using CTC Forced Alignment

Dayeon Ku, Hwa-Young Park, Hong Kook Kim*

*Gwangju Institute of Science and Technology, *AunionAI Co., Ltd.

ABSTRACT

This paper investigates a method to detect pauses in speech using Connectionist Temporal Classification (CTC) forced alignment, which is commonly employed in speech recognition for aligning speech and text. First, we align the input speech according to a given text, and delimit the word boundaries in the waveform domain. Then the between-word pauses are declared as pause intervals under duration constraint. The experimental results demonstrate that speech pause detection with CTC forced alignment maintains robust performance even in noisy environments, compared to a conventional method that relies on the root mean squared amplitude of speech with a predefined threshold.

Ⅰ. Introduction

One of the typical methods to detect pauses in speech signals is to, first, calculate root mean squared (RMS) amplitude in each frame, and then define an interval as a pause interval when it does not exceed a predefined threshold. However, this method has limitations due to its static threshold, failing to detect pause frames when background noise elevates their RMS amplitude above the predefined threshold. Consequently, such an RMS-based approach is not robust especially in noisy environments.

As an alternative, this paper employs Connectionist Temporal Classification (CTC) forced alignment for speech pause detection in noisy environments. CTC has been popularly used to perform a monotonic alignment between speech and text tokens in speech recognition [1]. Thus, CTC can estimate the start and end of each phoneme and/or word boundaries by finding the optimal paths using forward-backward computation [1]. Therefore, it is expected that CTC-based pause detection should yield more robust performance than RMS-based ones because CTC aligns phonemes and words boundaries based on probabilities, not a predefined threshold.

Ⅱ. Related Work

Energy-based approaches for pause detection typically rely on calculating the RMS energy of each speech frame [2][3]. In these methods, the average energy of the k-th frame whose length is N, E_k, calculated as $E_k = \sqrt{\frac{1}{N}\Sigma x_{k,i}{}^2}$, where $x_{k,i}$ is the i-th normalized speech sample at the k-th frame. A frame is classified as a pause if E_k falls below the predefined threshold. After classifying all the frames as either a pause or a speech frame, an pause interval is declared if the number of consecutive pause frames are greater than D [4]. A set of pause intervals in an utterance is denoted as $\boldsymbol{P}=(s_0, e_0),(s_1, e_1),\cdots,(s_{p-1}, e_{p-1})$ where p is the total number in this pause interval, $\boldsymbol{P}$, and s_i and e_i denote the starting and ending frames in the i-th pause frame of $\boldsymbol{P}$, respectively. Note that $(s_i - e_i) \geq D$ for all i. However, this approach has limitations due to its reliance on a fixed threshold, making it vulnerable to noisy environments where background noise can elevate energy levels during silent periods above the threshold.

Ⅲ. Method

This paper, we propose a noise-adaptive pause detection method using a CTC forced aligner to determine word boundaries. The CTC forced aligner utilizes a multilingual Wav2Vec2 model to determine word boundaries, outputting time-based markers represented by the start and end times for each word in seconds.

First of all, we pad short silent frames at the beginning and end to ensure the audio meets the minimum length required for reliable alignment without affecting its quality. Then, we apply the CTC aligner to speech input sampled at a rate of 22.05 kHz, where the frame-wise processing is carried out with a hop length of 256. Thus, the time-based markers associated with word boundaries are subsequently converted into speech frame indices to identify potential pause intervals. Next, to detect meaningful pauses from within-word or between-word short pause, the intervals that exceed 43 frames ($\cong$ 500ms) are classified as pauses. This is because the duration of a pause corresponding to a comma in text typically ranges from 380 to 670ms [5], and pauses shorter than 500ms are most frequently used by speakers [6]. This procedure finally generates a set of pause intervals

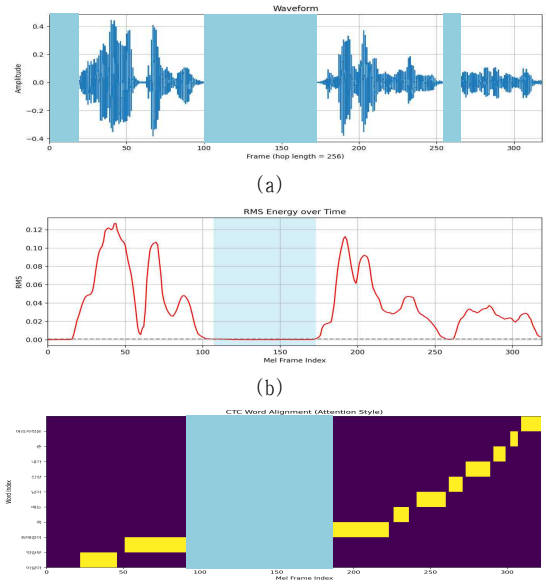

Figure 1: Illustration of pause detection results applied to (a) clean speech: (b) RMS-based and (c) proposed CTC-based detection.

$\boldsymbol{P}=(s_0,e_0),(s_1,e_1),\cdots,(s_{p-1},e_{p-1})$. Note that we only consider the pause intervals existing between the first and the last word in an utterance.

Fig. 1 compares the performance of the two pause detection methods on a source-separated Korean content audio sample. Fig. 1(a) is the waveform of the clean speech, and the boxes are the ground truth pauses located at $s_0=0$, $e_0=20$, $s_1=96$, $e_1=179$, $s_2=249$ and $e_2=266$. As mentioned earlier, (s_0,e_0) and (s_2,e_2) are placed before and after the main speech, so it should not be considered in the pause detection step. As displayed in Fig. 1, both the RMS-based and CTC-based methods were all able to identify the pause intervals, such as $(s_{rms}=107, e_{rms}=173$, and $s_{ctc}=91, e_{ctc}=185)$, compared with the ground truth. Note that the threshold was set to 0.001 and D=18 in the RMS-based pause detection.

Next, we mixed a noise with the signal-to-noise ratio (SNR) of 20 dB to the clean speech shown in Fig. 1(a). As illustrated in Fig. 2(b), RMS-based method failed to detect pauses because the RMS energy level during silent periods was above the set threshold. Conversely, the CTC-based approach was not significantly impacted by the noise and consistently provided pause detection performance $(s_{ctc}=91, e_{ctc}=185)$, as shown in Fig. 2(b).

Ⅳ. Conclusion

The inability to deliver noise-adaptive pause detection was a key limitation of the RMS-based approach. Based on the experiment performed in this paper, CTC-based approach has outperformed the RMS-base one with the predefined threshold in noisy environments. It was concluded in this paper that the proposed CTC-based pause detection was more robust especially in real-life applications such as dubbing, personal assistants, and real-time speech translation.

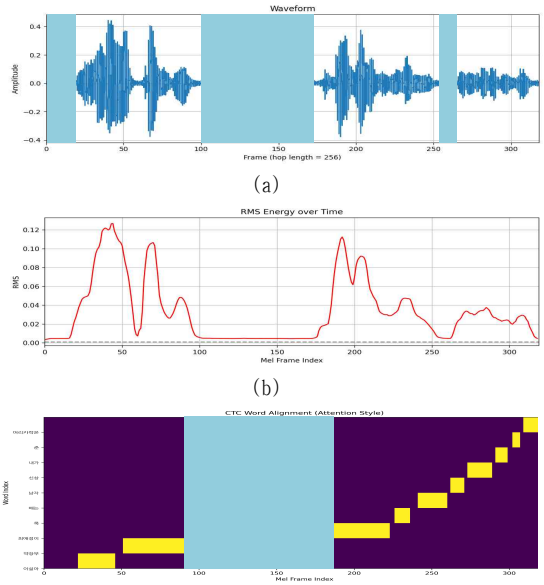

Figure 2: Illustration of pause detection results applied to (a) noisy speech of 20 dB SNR: (b) RMS-based and (c) proposed CTC-based detection.

ACKNOWLEDGMENT

This work was partly supported by the Technology development Program of MSS [RS-2025-21432982] and the program through the Innopolis Foundation funded by Ministry of Science and ICT [2022-DD-UP-0312].

Reference

[1] Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in Proceedings of the 23rd International Conference on Machine Learning, pp. 369-376, Pittsburgh, PA, USA, 2006.

[2] Özaydın, S. "Examination of energy based voice activity detection algorithms for noisy speech signals," European Journal of Science and Technology, vol. 17, pp. 442-449, 2019.

[3] Lesenfants, D., Vanthornhout, J., Verschueren, E., and Francart, T. "Cortical auditory responses index the contributions of different RMS-level-dependent segments to speech intelligibility," Journal of Neurophysiology, vol. 126, no. 2, pp. 463-473, 2021.

[4] De Jong, N. H., and Bosker, H. R. "Choosing a threshold for silent pauses to measure second language fluency," Proceedings of DiSS 2013, The 6th Workshop on Disfluency in Spontaneous Speech, pp. 17-20, Stockholm, Sweden, 2013.

[5] Yamashita, K., Liu, S., Sugino, T., O'Connell, D. C., and Kowal, S. "How pause duration influences impressions of English speech: Comparison between native and non-native Speakers," Frontiers in Psychology, vol. 13, pp. 1-15, 2022.

[6] Hoffer, J. Quantifying Speech Pause Durations in Typical English Speakers, Master's Thesis, Brigham Young University, 2023.