SpeechWorld at VLSP 2026 SQA: Simulation-Guided DSP–SSL Hybrids for Cross-Lingual Speech Quality Diagnosis

Tri-Nhan Do, Long-Hai Tran, Minh-Toan Tran

Independent Researchers, Vietnam

dotrinhan99@gmail.com, longhai7803@gmail.com, tranminhtoan140601@gmail.com

Abstract

The VLSP 2026 Speech Quality Assessment task asks for a quality score and the time-localised detection of noise, distortion and unnatural pauses. We propose a DSP–SSL hybrid that concatenates ten signal-processing frame features with a multilingual self-supervised encoder and is trained on QualiSpeech plus a Vietnamese degradation simulator. An ensemble of four w2v-BERT 2.0 hybrids and a WavLM-based model reaches a final score of 0.617 on the private test. The released labels reveal two shifts: a domain shift in language and fault characteristics, with short and often silent distortion events, and a shift in tag distribution, with unnatural pauses in only 2.6% of the private files against 34–44% in QualiSpeech.

Index Terms: speech quality assessment, fault localisation, data simulation, self-supervised models, Vietnamese

1. Introduction

A conventional speech quality assessment system returns one score; the VLSP 2026 SQA task additionally asks what is wrong and where. Systems output a quality score in [1,5] and, for each fault type, the time intervals where it occurs. The score is

$$\text{Final} = 0.4\,\text{PCC} + 0.3\,\text{F1}_{\text{macro}} + 0.3\,\text{IoU}_{\text{macro}}, \qquad (1)$$

where F1 is computed per fault at utterance level and IoU on a 20 ms grid over ground-truth-positive files, both averaged with equal weight over the three faults. Training, validation and public-test data come from QualiSpeech [1], an English corpus of which about half is TTS output; the 302-file private test consists of Vietnamese speech. The central difficulty is therefore a language and degradation shift that the English public test cannot measure.

Our contributions are: (i) a hybrid model that adds per-frame signal-processing (DSP) evidence to multilingual self-supervised (SSL) encoders, trained with a Vietnamese degradation simulator that provides exact fault intervals; (ii) an ensemble with English-only WavLM models and an analysis of why it helps; (iii) evidence that a simulated in-language dev set ranks systems for the private test better than the public test does; and (iv) a post-competition analysis of the released labels that separates a domain shift in fault characteristics from a shift in tag distribution and quantifies the error each causes, together with a step-wise ablation from the official training data to our best submission.

2. Related work

Non-intrusive quality prediction has moved from MOS regressors such as MOSNet [2] and NISQA [3], which also predicts quality dimensions, to DNSMOS [4] for noise suppressors and SSL-based predictors such as UTMOS [5], which led the VoiceMOS Challenge [6]. These systems output utterance-level scores. QualiSpeech [1] adds natural-language descriptions with fault types and time spans, which the VLSP task turns into a localisation problem. Multilingual SSL encoders such as w2v-BERT 2.0 [7], Whisper [8] and MMS [9] offer representations beyond English. Data simulation with additive noise [10] and room impulse responses [11] is standard for robust speech recognition; here we use it to create labelled fault intervals in a target language without any human annotation.

3. Data

3.1. QualiSpeech

We parse QualiSpeech's free-text fault descriptions (e.g. "background noise similar to whistle from 0.8s to 1.32") into presence flags and intervals, and mask ambiguous descriptions in the loss. This gives 10,558 training (9,987 with all three presence labels), 1,985 validation and 1,647 public-test files with complete labels. Table 1 summarises the labels; Sec. 6 compares them with the private test.

3.2. Simulated Vietnamese data

Clean Vietnamese speech (PhoAudioBook, Vivoice, LSVSC, ViSEC, Dolly) is degraded with exact fault intervals (Fig. 1a). Noise comes from Vietnamese music, traffic, in-car and TV recordings, MUSAN-style noise [10] and synthetic static. Distortion is produced by clipping, bit-crushing, aliasing, Griffin-Lim, pitch shifting, codec-like quantisation, musical noise, AGC pumping and clusters of short digital-zero dropouts. Unnatural pauses are produced by inserted gaps, stutters and choppy speech. The simulator was refined over several iterations towards more realistic degradations; Fig. 2 lists the operators of the final version, which generates 12k files (24.8 h, 1k held out as dev). Simulated files carry no MOS, so the quality loss is masked on them. As in QualiSpeech, inserted silences are labelled as pauses.

4. Proposed DSP–SSL hybrid

Our main system (Fig. 1) combines a fine-tuned self-supervised (SSL) encoder with a signal-processing (DSP) branch on the 20 ms grid of the task and is trained on QualiSpeech plus the simulated Vietnamese data.

4.1. SSL branch

The encoder (w2v-BERT 2.0 [7], WavLM-large [12], the Whisper-large-v3 encoder [8] with a 16 s window, or MMS-300m [9]) is fine-tuned end to end with SpecAugment [13]. A

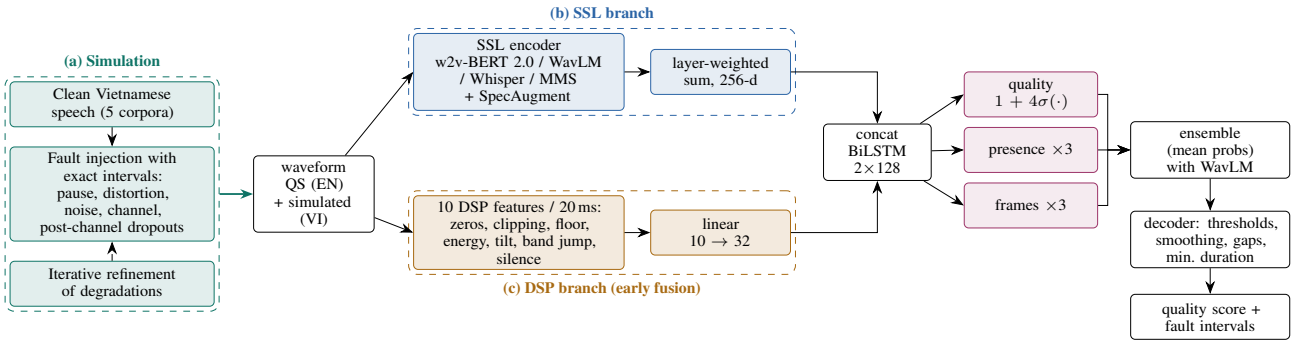


Figure 1: *Proposed DSP–SSL hybrid. (a) Clean Vietnamese speech is degraded with exact fault intervals; the degradations were refined over several iterations. (b, c) SSL frames are concatenated with DSP features before a BiLSTM (early fusion).*

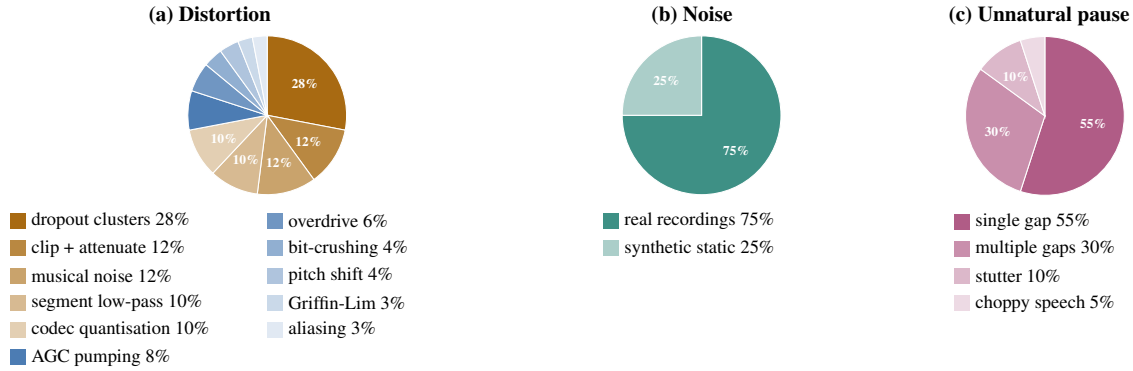


Figure 2: *Fault operators of the final simulator and their share among the injected faults of each type. 25% of the files are fault-free; each fault is otherwise injected with probability 0.40–0.45. Dropout clusters: 3–9 digital-zero gaps of 40–100 ms; clipping is followed by attenuation to a peak of 0.04–0.3; low-pass at 0.8–2 kHz. Noise: music, Vietnamese songs, traffic, in-car, babble, TV and rain recordings, SNR 10–28 dB. Pauses: gaps of 0.25–0.85 s filled with room tone; stutters repeat 80–250 ms chunks.*

Table 1: *Label statistics. Prevalence: share of files with each fault. Distortion length: median duration of a distortion interval; coverage: median share of a distortion-positive file covered by distortion.* †*Labels released by the organisers after the evaluation; used only for the analysis in Sec. 6.*

			Prevalence (%)			Distortion	
Set	Files	Hours	Noise	Dist.	Pause	length (s)	coverage
QS train	9,987	16.2	31.3	63.8	44.3	3.80	88%
QS dev	1,985	3.5	29.0	67.6	47.1	4.50	90%
QS public	1,647	2.5	33.4	56.7	33.7	4.00	90%
Simulated	12,000	24.8	38.3	38.3	34.0	2.03	29%
Private†	302	0.6	45.0	64.2	**2.6**	**0.56**	**11%**

softmax-weighted sum over all hidden layers is passed through LayerNorm, a 256-d linear layer, GELU and dropout, and resampled to the 20 ms grid.

4.2. DSP branch

SSL encoders see a level-normalised waveform and are trained on speech content, so digital artefacts leave weak traces: a run of exact zeros, a clipped plateau or a sudden loss of bandwidth. The DSP branch computes ten time-domain features per 20 ms frame from the raw waveform, before gain augmentation and cropping so that digital zeros and clipping are preserved: the share of digital-zero samples and membership in a zero run of at least 40 ms (dropouts), the share of clipped samples, frame energy, the local noise floor, spectral tilt and its deviation from the file's median (band-limiting, codecs), the energy change, and a silence flag with the length of the surrounding silent run (pauses). Levels are expressed relative to file statistics (the 5th and 95th percentiles of frame energy), which makes the features independent of recording level, speaker and language; all features are scaled to roughly unit range and clipped to $[-3, 3]$. A linear layer with GELU maps them to 32 dimensions.

4.3. Fusion, heads and loss

The 256-d SSL and 32-d DSP frames are concatenated (early fusion) and processed by a two-layer BiLSTM with 128 units per direction. A linear frame head gives three fault probabilities per frame. Attention pooling with four learned queries (one per fault, one for quality), concatenated with global max pooling, feeds three presence logits and a quality MLP with output $1 + 4\sigma(\cdot) \in [1, 5]$. The loss is

$$\mathcal{L} = \mathrm{MSE}_q + \mathrm{BCE}_{\mathrm{pres}} + \mathrm{BCE}_{\mathrm{frame}} + 0.5\,(1 - \mathrm{IoU}_{\mathrm{soft}}), \quad (2)$$

where the frame BCE is masked where interval labels are unknown, the soft IoU is averaged over fault-positive files, and the quality term is masked on simulated files, which carry no MOS. Random 12 s crops keep features and labels aligned; the presence of a cropped positive is re-derived from its intervals.

Table 2: *Official private-test results of our three main submissions, overall and per fault (best of five counts; two further variants scored 0.583 and 0.511).*

				Overall				Noise		Distortion		Pause	
#	System	Backbone	Data	Final	PCC	F1	IoU	F1	IoU	F1	IoU	F1	IoU
4	WavLM soup	WavLM-large	QS	0.551	0.734	0.535	0.323	0.772	0.612	0.782	0.242	0.051	**0.116**
3	DSP–SSL hybrid	w2v-BERT 2.0	QS + simulated VI	0.592	0.740	0.575	0.412	**0.870**	0.697	0.763	0.462	0.093	0.078
1	4 hybrids + WavLM soup	both	QS + simulated VI	**0.617**	**0.789**	**0.581**	**0.425**	0.844	**0.698**	**0.804**	**0.499**	**0.094**	0.079

4.4. Decoding and training

For each fault, the presence probability is thresholded first. For a positive file, the frame probabilities are smoothed by a moving average and thresholded, gaps up to g are closed and intervals shorter than d are dropped; if nothing survives, the region above half of the maximum probability is kept, so that presence and intervals stay consistent. Thresholds, smoothing, g and d are tuned per fault by coordinate search on 0.5 F1 + 0.5 IoU on dev data. Training uses AdamW (encoder 1e-5–2e-5, heads 1e-3), batch 8, bf16 and 5–7 epochs on one A100-40GB.

4.5. Model ensemble

The hybrids are combined with English-only WavLM models developed in parallel on QualiSpeech. The main one is WavLM-large [12] with a NISQA-style head [3] (layer sum, mel branch, auxiliary outputs), trained jointly on quality, frame BCE, presence BCE and a 0.2-weighted IoU loss, with waveform normalisation and a weight-averaged *soup* of epochs 3–6. Further WavLM variants came from an automated loop over losses (dense frame BCE, balanced IoU, quality ranking and contrastive terms), learning rates and decoder calibration. All models share the output format, so an ensemble simply averages the quality, presence and frame probabilities of its members and is decoded with one decoder tuned on dev data. Our final ensemble averages four w2v-BERT 2.0 hybrids, trained on different simulated data and seeds, with the WavLM soup.

5. Experiments and results

5.1. Model selection

Because the English public test cannot reveal Vietnamese behaviour, we re-scored every model of the team under one protocol before submission: the decoder is tuned on QualiSpeech dev and simulated dev sets of earlier simulator iterations, and evaluated on the public test and on the held-out dev set of the final simulator (*simulated dev set*). On the public test all single models lie within 0.663–0.684. The simulated dev set separates English-only models (IoU $\approx$0.50) from simulation-trained hybrids (up to 0.87); with comparable data, the multilingual w2v-BERT 2.0 is the strongest encoder, while Whisper, MMS and WavLM hybrids trail by 0.03–0.05 IoU. Sec. 6.4 checks this selection with the released labels.

5.2. Official results

Table 2 shows the role of the backbone and of the training data in our three main submissions. The English-only WavLM soup (#4) reaches 0.551. The w2v-BERT 2.0 hybrid trained with simulated Vietnamese data (#3) has a similar PCC but localises far better (IoU 0.323→0.412, distortion IoU 0.242→0.462, Table 2). The ensemble of both families (#1) keeps the localisation of the hybrids and raises PCC to 0.789, giving our best score of 0.617; Sec. 6.6 analyses why. Noise is detected well by all three systems, and pause detection is weak for every system (Sec. 6.3).

5.3. Computational cost

Each hybrid trains in 2–7 hours on one A100-40GB (5–7 epochs); generating the 12k simulated files takes a few minutes on 8 CPU cores. Inference on the 302 private files takes under a minute per model, so the five-model ensemble adds negligible cost.

6. Analysis

After the evaluation the organisers released the private-test labels. This section uses them; none of it informed our submissions.

6.1. Tag distribution shift

Table 1 shows that the tag distribution of the private test differs strongly from QualiSpeech: unnatural pauses almost disappear (2.6% of the files), noise is more frequent and quality labels are lower. The tag combinations change accordingly. QualiSpeech files often combine distortion with a pause or carry all three faults, whereas private files mostly carry distortion alone, noise alone or both, and none carries all three. Our decoders, calibrated on dev sets with a much higher pause prevalence, therefore start from a pause prior that is far too high.

6.2. Shift in fault characteristics

Besides the language, the faults themselves differ. We measure every labelled interval and every internal silent gap, i.e. at least 0.26 s of frames 25 dB below the file's speech level with speech on both sides.

Noise. Private noise is subtler: noise-labelled files are only slightly noisier than noise-free ones, much less so than in QualiSpeech, so a noise threshold learned on QualiSpeech does not transfer.

Distortion. QualiSpeech distortions mostly span the whole file, as expected for codec, TTS or processing artefacts. Private distortions are short local events, and almost half of them are silent gaps.

Silent gaps. QualiSpeech often labels internal silent gaps as pauses and almost never as local distortion. The private test leaves most gaps unlabelled as natural and labels the abnormal ones as distortion, almost never as pauses. The QualiSpeech behaviour follows from its sources: most of its pause labels come from NISQA files with simulated transmission impairments, where packet-loss gaps are labelled as pauses, whereas pause labels on TTS/VC files mark unnatural rhythm rather than silence. Our models inherit this reading: the pauses predicted by #1 fall mostly inside private distortion intervals, far above chance, on the gaps that the annotators marked as abnormal (Fig. 3). The model thus finds the abnormal gaps, but names them as QualiSpeech does.

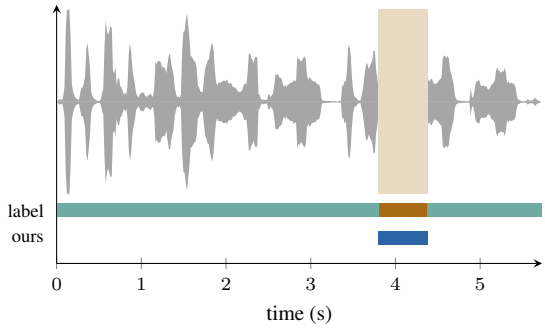


Figure 3: *Private file* `pt_00925`*. Label lane: noise over the whole file (green) and distortion (orange) on the silent gap at 3.80–4.38 s. Our submission #1 finds exactly this gap but labels it as an unnatural pause (blue), as QualiSpeech would.*

6.3. Error analysis

Pauses account for almost half of the gap between submission #1 and a perfect score. Post-hoc counterfactuals with the released labels attribute this loss to the two shifts above; submission #3 shows the same pattern.

Tag prior. #1 predicts a pause in 98 files, whereas only 8 are positive. Removing these false positives is the largest single gain.

Gap labelling. Relabelling the predicted pauses as distortion markedly improves distortion F1 and IoU.

The two shifts are linked: because abnormal gaps are labelled as distortion, pauses become rare. Fixing both would leave mainly PCC and noise as error. Both readings of a gap are defensible (packet loss is a channel distortion, a speaker's hesitation is a pause), so stating the convention in the task description, or releasing a small labelled in-domain dev set to calibrate fault priors, would help future editions.

6.4. Reliability of the development sets

The released labels allow a check of the model selection of Sec. 5.1 (Fig. 4). The simulated dev set ranks systems better than the public test, especially among single models, on which the public test is almost uninformative. It is not a precise guide, however: it measures localisation only, so English-only models with good quality prediction score higher on the private test than their IoU on the simulated dev set suggests. The public test shares sources and labelling with the training data and hides both shifts above; the simulated dev set could not reveal them either, since our simulator copied the QualiSpeech pause prior and labelling of silent gaps.

6.5. Step-wise ablation

To measure what each component contributes, we re-trained w2v-BERT 2.0 under controlled settings (Table 3). Three points stand out. First, models trained on QualiSpeech alone match the simulation-trained ones on the public test but not on the private test, even as a large ensemble; the gain comes from in-language simulated data, mainly through localisation. Second, DSP features help only together with simulated data that contains the corresponding faults. Third, the encoder matters as much as the data: a WavLM hybrid with the same data and recipe stays far below w2v-BERT.

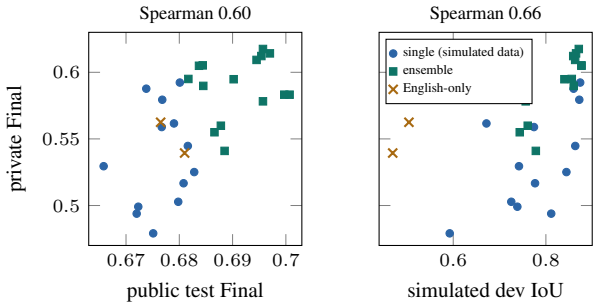


Figure 4: *Private Final (released labels) against the public test Final and the IoU on the held-out simulated dev set for 30 systems, each with its pre-submission decoder.*

Table 3: *Step-wise ablation on the private test. Each "+" row adds to the row above; the QS-only and DSP rows average two seeds. Bottom: an ensemble of 12 QualiSpeech-only models of the team, and a WavLM-large hybrid with the same data and recipe.*

System (w2v-BERT 2.0)	Private	PCC	F1	IoU
QS only	0.542	0.688	0.545	0.343
+ simulated VI data	0.564	0.677	0.570	0.409
+ DSP features	0.586	0.732	0.569	0.408
+ ensemble of 4 hybrids	0.605	0.757	**0.581**	**0.427**
+ WavLM soup (#1)	**0.617**	**0.789**	**0.581**	0.425
QS-only ensemble (12 models)	0.571	0.764	0.519	0.367
WavLM hybrid	0.494	0.587	0.527	0.338

6.6. Ensemble analysis

Averaging helps as far as the members' errors are not shared. The four hybrids share the encoder, the head and most of the training data, so their errors are strongly correlated and averaging them helps only moderately. The WavLM soup differs in backbone, head and training data; its errors are largely independent of those of the hybrids, so adding it improves both the quality correlation and distortion localisation, although on its own it is no better than a hybrid. The gain of the ensemble therefore comes from diversity in backbone and training data rather than from the number of members.

7. Conclusion

We presented a DSP–SSL hybrid for joint speech quality prediction and fault localisation, trained on QualiSpeech and a Vietnamese degradation simulator with exact fault intervals. An ensemble of four w2v-BERT 2.0 hybrids and a WavLM-based model reached a final score of 0.617 on the private test. The ablation shows that in-language simulated data and a multilingual encoder drive most of the gain, that DSP features help once the simulated data contain the faults they expose, and that the ensemble profits from diverse rather than many similar members. The English public test could not separate our models, whereas a small simulated in-language dev set ranked them better. The released labels reveal a shift in fault characteristics, with short and often silent distortions, and a shift in tag distribution, with rare pauses; together they account for most of the pause error. Natural next steps are to extend the simulator towards such local distortions, to calibrate fault priors for the target domain, and to align label conventions between training and test data.

8. References

[1] S. Wang, W. Yu, X. Chen, X. Tian, J. Zhang, L. Lu, Y. Tsao, J. Yamagishi, Y. Wang, and C. Zhang, "QualiSpeech: A speech quality assessment dataset with natural language reasoning and descriptions," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

[2] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "MOSNet: Deep learning-based objective assessment for voice conversion," in *Proc. Interspeech*, 2019.

[3] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," in *Proc. Interspeech*, 2021.

[4] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. ICASSP*, 2022.

[5] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022," in *Proc. Interspeech*, 2022.

[6] W.-C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS challenge 2022," in *Proc. Interspeech*, 2022.

[7] Seamless Communication, L. Barrault *et al.*, "Seamless: Multilingual expressive and streaming speech translation," *arXiv preprint arXiv:2312.05187*, 2023.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[9] V. Pratap, A. Tjandra, B. Shi *et al.*, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.

[10] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[11] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. ICASSP*, 2017.

[12] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[13] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019.