Odyssey 2026
23-26 June 2026, Lisbon, Portugal

Speech Quality Embeddings for Improved Detection and Classification of Degradations in Speech Signals

Michael Kuhlmann[1], *Tobias Cord-Landwehr*[1], *Reinhold Haeb-Umbach*[1]

[1]Paderborn University, Germany

{kuhlmann,haeb}@nt.uni-paderborn.de

Abstract

Automatic subjective speech quality assessment (SSQA) traditionally estimates speech quality on an utterance or system level. While this resolution was adequate for older transmission or synthesis systems that produced speech signals of mediocre quality, modern systems generate high-quality speech with degradations that may occur only locally. With suitable model architectures and regularization losses, SSQA models trained with utterance-level targets can also yield useful local predictions of speech quality. In this work, we extend such models to produce frame-level embeddings that cluster by degradation type. Specifically, we employ a partial mix-up strategy on a parallel corpus of clean and degraded utterances and apply a contrastive loss to distinguish between degradation types. Through experiments on both in- and out-of-domain data, we demonstrate that our approach improves degradation detection and enables the identification of degradation types by analyzing embedding clusters.

Index Terms: speech quality, quality embeddings, degradation detection

1. Introduction

Automatic subjective speech quality assessment (SSQA) assigns a quality score to speech signals that reflects subjective quality perception (e.g., poor or excellent) [1]. Usually, these approaches work *non-intrusively*, i.e., without knowledge of a clean matching reference signal. Furthermore, they are trained with mean opinion score (MOS) supervision - gathered from listening experiments - and provide a single score for the complete utterance (*global* score). However, the expressivity of a global score is limited and, for example, cannot inform the user about temporally constrained artifacts in the speech [2, 3]. Supplementing global evaluation with frame-level (*local*) evaluation can help with artifact localization [3], increase interpretability [4, 2, 3], and may be used as a learning signal to reduce such artifacts in future model iterations [5, 6]. Beyond localizing artifacts, automatic identification of artifact types can further increase interpretability and retrieval of specific artifacts [7].

Only a few prior works have looked into extending speech quality assessment (SQA)[1] models to include these capabilities. Quality-Net [4] trained an SQA model with frame-level regularization to predict plausible frame-level quality scores. Recently, frame-level regularization was applied to stronger SSQA models based on self-supervised learning (SSL) [2, 3]. To detect local degradations, they tuned and applied a threshold on the predicted frame-level MOS (*MOS-based degradation detection*). Because no training-scale data with frame-level quality scores exist, all these approaches rely on utterance-level targets from which frame-level quality scores should be learned.

Learning to detect regions of signal degradation from a global, i.e., *weak*, target shows large resemblances to the research field of sound event detection (SED) [8, 9, 10]. Recently, frame-level and time-aligned audio-language models [11, 12] were developed that replace the static set of known classes commonly used in many SED approaches with textual descriptions. To do so, these approaches share two key factors: training with frame-level (*strong*) (pseudo-)targets and using a contrastive loss to align frame-level embeddings with text annotations.

On the frontier of identifying degradations, Cumlin et al. [7] found that the latent space of DNSMOS-like [13] models clusters by degradation type. They used a kNN classifier to identify known impairments and could further improve the results by training on impairment-augmented data with PESQ [14] or ViSQOL [15] labels. However, their observation holds for utterance-averaged embeddings, in which impairments dominate the speech signal, and it is unclear whether it generalizes to local degradations. NOMAD [16] used a triplet loss between degraded signals to learn perceptual embeddings. As positive pairs were constructed using close NSIM [17], the embeddings capture degradation intensity rather than type.

In this work, we show that the non-intrusive *joint detection and identification* of local degradations in a speech signal can be learned by borrowing concepts from frame-wise audio language models, namely using data augmentation to generate strong *pseudo-targets* [11] and using a contrastive loss to align similar concepts in the embedding space [11, 12]. Our contributions are twofold. First, using a parallel data corpus of clean reference and degraded speech signals with utterance-level MOS-scores, we propose a *partial mix-up* strategy to generate training data with local degradations. Using frame-level pseudo-targets from a pretrained SSQA model [3], we train a new frame-level SSQA model on these mix-up data. Compared to the baseline [3], MOS-based degradation detection can be significantly improved when adding this frame-level supervision during training. Second, to improve the similarity of embeddings derived from similarly degraded frames, we build upon the partial mix-up augmentation strategy and add a frame-level supervised contrastive loss [18], exploiting knowledge of the degradation classes used during mix-up. We find that switching to an embedding-based degradation detection method, using a clean reference embedding as the enrollment, yields the best detection performance. Finally, an analysis of the latent embedding space reveals that different degradation types form distinct clusters, but that cluster purity degrades under multiple simultaneous degradations or when switching to out-of-domain degradations.

[1]Following [1], we use the term speech quality assessment to denote the broader class of models that are not only trained with human preference data but with signal-based intrusively-derived labels like PESQ.

10.21437/Odyssey.2026-37

2. Local speech quality assessment

Local subjective speech quality assessment (LSSQA)[2] aims to estimate quality at a finer resolution than the utterance level. Similar to utterance-level assessment, encoder-decoder models [1] can be used to infer frame-level quality scores[3] [2].

Given a dataset $\mathcal{D} = \{(s_1(t), y_1), (s_2(t), y_2), \dots\}$ of speech signals $s_i(t)$ with utterance-level MOS annotation y_i, the frame-level scores are inferred from an audio signal as follows. First, $s(t)$ is processed by an encoder, which extracts frame-level embeddings $\mathbf{X} = \mathrm{Enc}(s(t)) = [\mathbf{x}_1, \dots, \mathbf{x}_l, \dots, \mathbf{x}_L]$ with frame index l. Then, a decoder network is applied to the frame-wise audio embeddings, estimating the MOS of each frame as $\hat{\mathbf{q}} = \mathrm{Dec}^{\text{MOS}}(\mathbf{X}) = [\hat{q}_1, \dots, \hat{q}_l, \dots, \hat{q}_L]$. Since frame-level targets are usually unavailable during training time, approaches rely only on utterance-level targets [19, 20, 21], make specific architectural choices [2], and add regularization losses during training to infer frame-level scores [4, 3]. In [2], a conventional SSQA loss [22, 21] with a pool-last strategy was used

$$\mathcal{L}_{\text{SSQA}} = \frac{1}{B}\sum_{b=1}^{B} \max\left(\left|y_b - \frac{1}{L}\sum_{l=1}^{L}\hat{q}_{b,l}\right| - \delta, 0\right) + \mathcal{L}^{\text{MOS-con}}, \tag{1}$$

where y_b and $\hat{q}_{b,l}$ denote the target MOS and estimated l-th frame MOS of the b-th example in a mini-batch of size B, δ is a clipping parameter [22] and $\mathcal{L}^{\text{MOS-con}}$ is a contrastive loss on the utterance-level scores [21]. In [3], an additional consistency loss between data slices was added to reduce the influence of far-away context frames on the embeddings and scores

$$\mathcal{L}_{\text{LSSQA}} = \mathcal{L}_{\text{SSQA}} + \frac{1}{\Delta L}\left(\sum_{l=l_0}^{l_0+\Delta L} \|\mathbf{x}_l - \mathbf{x}_l^{\text{slice}}\|_2^2 + |\hat{q}_l - \hat{q}_l^{\text{slice}}|\right), \tag{2}$$

where $\mathbf{x}^{\text{slice}}$ and $\hat{q}^{\text{slice}}$ denote encoder embeddings and frame-level scores that are extracted from an input signal slice $s^{\text{slice}}(t)$ of duration ΔL frames (see [3] for details).

3. Improved frame-level scores

Local degradation detection is challenging because only utterance-level, i.e., weak, quality targets are available at scale. As motivated at the beginning, we expect that training with frame-level targets will improve detection performance. To this end, we propose a straightforward data augmentation strategy to produce frame-level labels for training an LSSQA model by defining pseudo-targets. These pseudo-targets are defined using a corpus of parallel data comprising clean reference and degraded speech utterances, $\mathcal{D}_{\text{ref}}$ and $\mathcal{D}_{\text{deg}}$, respectively[4]. We employ a pretrained LSSQA model [3] to obtain parallel sets of frame-level scores, $\mathcal{Q}_{\text{ref}}$ and $\mathcal{Q}_{\text{deg}}$, from $\mathcal{D}_{\text{ref}}$ and $\mathcal{D}_{\text{deg}}$, respectively. During training, we sample for each audio with probability p_{mixup} binary mix-up masks $m(t)$ and $\mathbf{m}_q = [m_1, \dots, m_L]$, with $m(t), m_l \in \{0, 1\}$, to generate augmented training data with frame-level pseudo targets

$$s^{\text{pseudo}}(t) = m(t)s^{\text{deg}}(t) + (1 - m(t))s^{\text{ref}}(t), \tag{3}$$

$$\mathbf{q}^{\text{pseudo}} = \mathbf{m}_q \odot \hat{\mathbf{q}}^{\text{deg}} + (1 - \mathbf{m}_q) \odot \hat{\mathbf{q}}^{\text{ref}}, \tag{4}$$

where $m(t)$ and $\mathbf{m}_q$ are defined in sample and frame resolution, respectively, and otherwise mask the same time points, and $\odot$ denotes element-wise multiplication. Importantly, for this to work, we require that the degradations persist throughout the complete utterance taken from $\mathcal{D}_{\text{deg}}$ with a similar strength. Then, it suffices that the pretrained model can predict frame-level scores and has a high utterance-level correlation with the target MOS; it need not be fine-tuned for detection. We then train a new LSSQA model with additional frame-level supervision based on the pseudolabels $\mathbf{q}^{\text{pseudo}}$

$$\mathcal{L}_{\text{LSSQA}}^{\text{sup}} = \mathcal{L}_{\text{LSSQA}} + \frac{1}{BL}\sum_{b=1}^{B}\sum_{l=1}^{L}|\hat{q}_{b,l} - q_{b,l}^{\text{pseudo}}|, \tag{5}$$

where $\hat{\mathbf{q}}$ is now inferred as $\mathrm{Dec}^{\text{MOS}}\left(\mathrm{Enc}\left(\mathbf{s}^{\text{pseudo}}\right)\right)$ and we set $y_b = \frac{1}{L}\sum_{l=1}^{L} q_{b,l}^{\text{pseudo}}$ in $\mathcal{L}_{\text{SSQA}}$ when partial mix-up was applied.

4. Speech quality embeddings

For downstream applications, not only the location of a degradation in a signal, but also its *type* is relevant. This can be beneficial in system analysis, e.g., identifying frequent degradation types or retrieving specific degradations from large databases. Cumlin et al. [7] studied the latent space of DNSMOS-like [13] models and found that the utterance-level embeddings form clusters by degradation type. However, unlike DNSMOS, we perform time pooling at the decoder output, and we find that neither our encoder nor decoder embeddings cluster well by degradation type. Moreover, Deng [23] recently showed that audio embeddings do not always form clear clusters under different degradations and degradation strengths.

4.1. Contrastive learning for quality embeddings

To learn discriminative frame-level degradation-type embeddings, we assume that the degraded dataset used for pseudo-supervised training contains annotations about the degradation type c: $\mathcal{D}_{\text{deg}} = \{(s^{\text{deg}}(t), y, c)_i\}$ [20, 7]. Using these annotations, we add a supervised contrastive loss [18] to distinguish between degradation types as follows. After applying partial mix-up, a class label $c_l \in \Omega = \{1, \dots, K+1\}$ is obtained for each frame, where K is the total number of degradation classes during training, and an additional class is added to denote clean, non-degraded frames. To decouple MOS estimation from representation learning, a second decoder is added to extract frame-level bottleneck embeddings $\mathbf{Z}^{\text{scl}} = \widetilde{\mathrm{Dec}}^{\text{scl}}(\mathbf{X}) = [\mathbf{z}_1^{\text{scl}}, \dots, \mathbf{z}_L^{\text{scl}}]$, $\mathbf{z}_l^{\text{scl}} \in \mathbb{R}^{D_Z}$. Following [18], an additional linear projection $\tilde{\mathbf{z}} = \mathrm{Proj}^{\text{scl}}\left(\mathbf{z}^{\text{scl}}\right) \in \mathbb{R}^{D_P}$ is applied, where both $\mathbf{z}^{\text{scl}}$ and $\tilde{\mathbf{z}}$ are normalized to lie on the unit hypersphere. Figure 1 shows a schema of the full system.

Given a mini-batch of embeddings $\tilde{\mathcal{Z}}_b \in \mathbb{R}^{B\times L\times D_P}$ and corresponding frame-level class affiliations $\mathcal{C}_b \in \Omega^{B\times L}$, the following supervised contrastive loss is added during training:

$$\mathcal{L}^{\text{scl}} = -\frac{1}{BL}\sum_{b=1}^{B}\sum_{l=1}^{L}\mathcal{L}_{b,l}^{\text{scl}} \tag{6}$$

with

$$\mathcal{L}_{b,l}^{\text{scl}} = \frac{1}{|I_P(c_{b,l})|}\sum_{\substack{(b',l')\in I_P(c_{b,l})\\(b',l')\neq(b,l)}} \log\frac{\exp\left(\tilde{\mathbf{z}}_{b,l}^{\mathsf{T}}\tilde{\mathbf{z}}_{b',l'}/\tau\right)}{\sum_{\tilde{b}}\sum_{\substack{\tilde{l}\\(\tilde{b},\tilde{l})\neq(b,l)}}\exp\left(\tilde{\mathbf{z}}_{b,l}^{\mathsf{T}}\tilde{\mathbf{z}}_{\tilde{b},\tilde{l}}/\tau\right)}, \tag{7}$$

[2] In the following, SQA/SSQA always refer to the *automatic* quality assessment of speech.

[3] In the following, we will refer to these as *frame-level scores*.

[4] Such corpora can be obtained by simulating signal distortions [7, 20], voice conversion, or forced-aligned text-to-speech models.

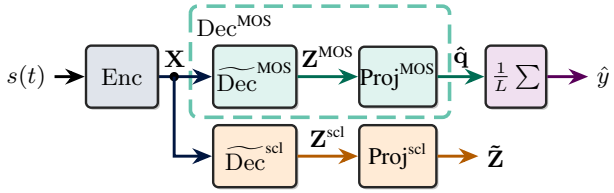

Figure 1: *Block schema of the full proposed model. The encoder* Enc *feeds two decoder heads:* Dec^{MOS} *for frame-level scores* $\mathbf{q}$*, followed by mean pooling for utterance-level estimates* $\hat{y}$*, and* $\widetilde{\text{Dec}}^{scl}$ *for frame-level embeddings* $\mathbf{Z}^{scl}$*, followed by a projection layer for contrastive training.* $\mathbf{Z}^{MOS}$ *denotes the embeddings of the MOS decoder before projection to the frame-level scores.*

where $I_P(c_{b,l}) = \{(b', l') \in \{1, \dots, B\} \times \{1, \dots, L\} : c_{b,l} = c_{b',l'}\}$ is the set of indices in the mini-batch that have the same class label $c_{b,l}$ and $\tau > 0$ is a temperature parameter. Essentially, this loss computes the cross-entropy between a softmax distribution of $B \cdot L$ elements over $B \cdot L$ classes and the multi-hot target distribution matrix $\mathbf{P}$, where $p_{bl,b'l'} = 1/|I_P(c_{b,l})|$ if $c_{b,l} = c_{b',l'}$ and 0 else [24]. Different from [18], the contrastive loss is applied to frame-level embeddings here, meaning that embeddings $\tilde{\mathbf{z}}_{b,l}$ and $\tilde{\mathbf{z}}_{b,l+\lambda}$ for some small value λ exhibit strong correlation. We consider the embeddings $\{(\tilde{\mathbf{z}}_{b,l_1}, \tilde{\mathbf{z}}_{b,l_2}) : |l_1 - l_2| < \lambda\}$ as a special case of self-contrast and exclude these pairs from the summations in eq. (7), where λ is a hyperparameter. Incorporating the supervised contrastive loss into the training of our LSSQA model, the total loss becomes

$$\mathcal{L}_{\text{total}} = \mathcal{L}^{\text{sup}}_{\text{LSSQA}} + \tau \mathcal{L}^{\text{scl}}, \tag{8}$$

where we follow [18] and scale $\mathcal{L}^{\text{scl}}$ with the temperature.

4.2. Handling non-degraded frames

With the partial mix-up strategy, many clean, i.e., non-degraded frames, will be involved in the computation of $\mathcal{L}^{\text{scl}}$. There are two possible ways to handle these frames: i) Treat non-degraded frames as a distinct class or ii) ignore their contribution to $\mathcal{L}^{\text{scl}}$. In the latter case, we modify eq. (6) to exclude specific classes

$$\mathcal{L}^{\text{scl}} = -\frac{1}{BL} \sum_{b=1}^{B} \sum_{l=1}^{L} \mathbb{1}\left(c_{b,l} \notin \overline{\Omega}\right) \mathcal{L}^{\text{scl}}_{b,l}, \tag{9}$$

where $\overline{\Omega}$ is the set of class indices that should be excluded. Note that the summations in the denominator in eq. (7) still run over the excluded classes. Especially in a sparse scenario, where the number of degraded frames is significantly lower than the number of clean frames, the latter case may help learn more distinct degradation clusters.

5. Local degradation detection: MOS-based versus embedding-based

Given estimated frame-level scores $\hat{\mathbf{q}}$, *MOS-based* detection of local degradations infers, for each frame, whether it suffers from a quality degradation by comparing its frame-level score to a score threshold q_{deg} that is tuned on a validation set.

Assuming that the embeddings form clusters by degradation type and that non-degraded frames cluster together, we can perform degradation detection with the encoder ($\mathbf{X}$) or decoder ($\mathbf{Z}^{\text{MOS}}$, $\mathbf{Z}^{\text{scl}}$, or $\tilde{\mathbf{Z}}$) *embeddings* as an alternative to MOS-based detection. Given a clean reference embedding for enrollment, we compute the cosine similarity between the enrollment embedding and all input embeddings, similar to NOMAD [16], and compare these similarities against a threshold to identify degraded frames. Compared to MOS-based detection, the detection performance depends on the choice of the enrollment and provides an easy way to tune the model to the inference data by selecting a matching enrollment. Note that a similar strategy can be used to retrieve frames exhibiting a specific degradation when a corresponding enrollment is available, analogous to a speaker verification [25, 26].

6. Experiments

To demonstrate that our contributions improve the detection of local degradations and the identification of degradation types, we conduct three evaluations. First, the proposed LSSQA extensions are evaluated with respect to their MOS- and embedding-based degradation detection. Then, the speech quality embeddings are assessed for their specificity in distinguishing among degradation types. Finally, the models' performance for joint degradation detection and degradation-type clustering is investigated.

6.1. Training & evaluation setup

We train all models on NISQA [20] and BVCC [27]. For NISQA, we apply the partial mix-up strategy as explained in Section 3. When training with $\mathcal{L}^{\text{sup}}_{\text{LSSQA}}$, we sample a mask $\mathbf{m}$ by sampling between 1 and 3 mask segments, each of duration between 200 ms and 1 s. When training with $\mathcal{L}_{\text{total}}$, we impose additional constraints during mask sampling so that the mix-up regions contain perceptible degradations almost surely: We exclude any utterances from mix-up where the utterance-level score of the reference $y^{\text{ref}} < 3.5$ or where the score of the degradation $y^{\text{deg}} > 4$, we only sample from speech regions[5], and we only sample from regions where $\mathbf{q}^{\text{ref}} - \mathbf{q}^{\text{deg}} > 0.5$. To construct the class labels, we read the metadata from the `NISQA_TRAIN_SIM` and `NISQA_VAL_SIM` subsets. In NISQA, often multiple degradations were applied to a single audio file. We treat the combination of multiple simultaneous degradations as a distinct class, as in [7]. This yields $K = 899$ distinct classes for the training split and $K = 371$ for the validation split, each split using the same 19 unique single degradations. All data preparation steps are open-sourced[6].

6.1.1. Validation and test data

As there exists no established benchmark dataset for local degradation detection, we use two synthetic datasets to report our results, each with a validation and test split.

`NISQA_VAL_SIM`-partial-mixup and **`NISQA_TEST_SIM`-partial-mixup**. These are in-domain validation and test datasets, as they include a similar set of degradations to the training data, and we use the same partial mix-up strategy as in the training to generate them, where `NISQA_TEST_SIM` is the combination of `NISQA_TEST_FOR` and `NISQA_TEST_P501`. For both splits, we use slightly longer degradation segments (between 400 ms and 2 s) than seen during training. The test split contains 9 unique single degradations and a total of $K = 36$ degradation classes.

LibriAugmented-partial-mixup. To test generalization to out-of-domain degradations, we apply partial mix-up to the *dev-clean* and *test-clean* splits of LibriAugmented [7] - an augmented version of LibriSpeech [29] with impairments applied from the Audiomentations library[7]. Since this dataset is not

[5] We use rVAD [28] to detect speech regions.

[6] `https://github.com/fgnt/local_sqa`

[7] `https://github.com/iver56/audiomentations/`

publicly available, we created our own version[8] with two modifications. We use background noises from CHiME-3 [30], and we replace the `GainTransition` impairment with the `Gain` impairment, which applies a constant gain to the full utterance and is better suited for our partial mix-up augmentation. Each utterance is impaired exactly twice: Once with a single impairment - choosing from 9 unique impairments - and once by applying two impairments - choosing from 6 impairment combinations - totalling $K = 15$ unique impairment classes.

6.1.2. Model configurations

For the encoder, we use a pretrained wav2vec2-large [31] that yields 1024-dimensional embeddings at a frame rate of 50 Hz and that is fine-tuned alongside the decoder. The MOS decoder is a 3-layer CNN with hidden layer sizes 256, 256, and $D_Z = 64$, kernel sizes[9] 11, 7, and 5, LeakyReLU activation, and batch normalization, followed by a linear projection to a scalar. For $\widetilde{\text{Dec}}^{\text{scl}}$, we use the same configuration with a projection to $D_P = 128$-dimensional embeddings, where the supervised contrastive loss is applied. For $\mathcal{L}_{\text{LSSQA}}$, we use the hyperparameters from [3]. Following [18], we use $\tau = 0.1$ in the computation of the supervised contrastive loss, and we set $\lambda = 10$, which equals half the size of the receptive field. Table 1 shows an overview of the trained models. All models are trained for 100 epochs with a batch size of 64 and an initial learning rate of 1×10^{-4} which is linearly decayed to 1×10^{-6}.

Table 1: *Training setup. When training with $\mathcal{L}_{total}$, we consider excluding clean frames from the set of positive classes (CON2).*

Model ID	Loss	p_{mixup}	Include clean
MOS embeddings			
Baseline [3]	$\mathcal{L}_{\text{LSSQA}}$ (2)	0.0	n/a
SUP1	$\mathcal{L}^{\text{sup}}_{\text{LSSQA}}$ (5)	0.5	n/a
SUP2	$\mathcal{L}^{\text{sup}}_{\text{LSSQA}}$ (5)	1.0	n/a
Contrastive embeddings			
CON1	$\mathcal{L}_{\text{total}}$ (6)	1.0	✓
CON2	$\mathcal{L}_{\text{total}}$ (6)	1.0	✗

6.1.3. Metrics

Local degradation detection. For embedding-based evaluation, we treat the similarity between the enrollment and each embedding as a single detection outcome, and report the frame-wise equal error rate (EER) (lower is better), and the minimum detection cost function (minDCF) [33], which is a weighted sum of the miss and false-alarm rates (lower is better). For minDCF calculation, we set $p_{\text{target}} = 0.01$, which is the assumed a priori probability of a degradation in our case. As per-frame detection does not assess whether the detections form connected regions that have a significant intersection with the ground-truth annotation, we report the area under curve of an intersection-based detection (I-AUC) [34] for both MOS- and embedding-based detection (higher is better). Intersection-based detection needs two hyperparameters: The detection tolerance criterion ρ_{DTC} and the ground truth intersection criterion ρ_{GTC}. We use $\rho_{\text{DTC}} = \rho_{\text{GTC}} = 0.7$, which is a common choice for sound event detection [35].

[8] Reproducible from `https://github.com/fgnt/frame-level-mos`

[9] The total receptive field is 400 ms, which matches STOI's frame duration [32].

Embedding analysis. We perform a *degradation verification* to check whether embeddings of the same degradation type are grouped together and distinguishable from other degradation types in the latent space. We compute the pairwise distances between the embeddings and report the EER of a detection using the distances to decide on the same class affiliation. In addition to the verification task, we perform retrieval by extracting a prototype for each degradation class from the validation split and computing distances between all test embeddings and the prototypes. For each test embedding, we take the prototype with the minimum distance and report the accuracy as the ratio of correctly selected prototypes.

Joint detection and assignment. Here, we perform an agglomerative clustering, assuming the number of degradation classes is known. Each unique combination of degradations is handled as an individual class, and for `NISQA_TEST_SIM`-partial-mixup, all regions with more than two simultaneously active degradations are omitted from clustering to keep the number of clusters tractable. Then, the adjusted rand index (ARI) in the range −1 to 1 (higher is better) is evaluated to assess the cluster purity compensated against a random assignment.

6.2. Local degradation detection

We perform MOS- and embedding-based detection as described in Section 5. Note that our metrics are threshold-independent, so we do not tune any thresholds on the validation split. For embedding-based detection, we use the validation splits as the enrollment set. To compute the enrollment embedding, we compute an utterance embedding as the mean of the frame embeddings and average all utterance embeddings.

Table 2 shows the results for the in-domain test data. Adding frame-level pseudo labels to the training significantly improves the performance of MOS-based detection (Baseline I-AUC: 0.01, SUP2 I-AUC: 0.52). Higher mix-up ratios lead to better results (SUP1 I-AUC: 0.25, SUP2 I-AUC: 0.52). The embedding-based detection for the MOS-only models also profits from the supervision (Baseline $\mathbf{x}$-I-AUC: 0.04 versus SUP2 $\mathbf{x}$-I-AUC: 0.57). Note that in SED, an I-AUC of around 0.6 is considered state-of-the-art [36]. However, detection costs are high with the lowest minDCF of 0.87 achieved by SUP1.

Training with the supervised contrastive loss leads to a significant drop in EER (SUP2: 11.1%, CON1: 3.87%) and a significant increase in I-AUC for embedding-based detection (SUP2: 0.58, CON1: 0.91). Including clean frames as a distinct positive class yields a significant performance improvement compared to excluding them (e.g., CON1 minDCF: 0.60, CON2 minDCF: 0.84). Regarding the embedding choice, the decoder bottleneck embeddings $\mathbf{z}^{\text{scl}}$ perform most robustly. The supervised contrastive loss slightly improves MOS-based detection performance (CON2 best MOS-I-AUC of 0.67); however, a large gap to embedding-based detection remains (CON1 best embedding-I-AUC of 0.91); hence, embedding-based detection should be preferred. Figure 2 shows a comparison between MOS- and embedding-based detection on a single example. The better performance of the embedding-based detection stems from a larger margin to the threshold for both clean and degraded frames, and from higher sensitivity to degradation frame onsets (here for the first degraded segment).

Table 3 shows the result on out-of-domain degradations. We observe the same trends and improvements as in the in-domain case, with slightly better MOS-based detection (CON2-I-AUC: 0.78 versus 0.67), but slightly worse EER (CON1 $\mathbf{z}^{\text{scl}}$-EER: 4.88% versus 3.87%).

Table 2: *Embedding- vs. MOS-based detection of degraded frames.* $I_{0.7,0.7}$*-AUC denotes the AUC of an intersection-based detection with* $\rho_{DTC} = \rho_{GTC} = 0.7$*. Best results for each metric and embedding type are shown in* ***bold****. Overall best results per metric are additionally* ***underlined****. Testset:* `NISQA_TEST_SIM`*-partial-mixup. Enrollment set:* `NISQA_VAL_SIM/ref`*.*

	Model ID	Embedding-based									MOS-based
		Frame-EER [%]			Frame-minDCF			$I_{0.7,0.7}$-AUC			$I_{0.7,0.7}$-AUC
without $\mathcal{L}^{scl}$		$\mathbf{x}$	$\mathbf{z}^{MOS}$		$\mathbf{x}$	$\mathbf{z}^{MOS}$		$\mathbf{x}$	$\mathbf{z}^{MOS}$		
	Baseline [3]	19.21	17.06		1.00	1.00		0.04	0.02		0.01
	SUP1	15.36	13.34		0.99	**0.87**		0.47	0.48		0.25
	SUP2	12.71	**11.01**		1.00	1.00		0.57	**0.58**		0.52
with $\mathcal{L}^{scl}$		$\mathbf{x}$	$\mathbf{z}^{scl}$	$\tilde{\mathbf{z}}$	$\mathbf{x}$	$\mathbf{z}^{scl}$	$\tilde{\mathbf{z}}$	$\mathbf{x}$	$\mathbf{z}^{scl}$	$\tilde{\mathbf{z}}$	
	CON1	**5.22**	**3.87**	**3.93**	**0.97**	**0.60**	**0.60**	**0.86**	**0.91**	**0.91**	0.65
	CON2	9.35	5.39	33.8	0.99	0.84	1.00	0.50	0.81	0.10	**0.67**

Table 3: *Embedding- vs. MOS-based detection on unseen degradations. Best results for each metric and embedding type are shown in* ***bold****. Overall best results per metric are additionally* ***underlined****. Testset: LibriAugmented/test-clean-partial-mixup. Enrollment set: LibriTTS/dev-clean.*

	Model ID	Embedding-based									MOS-based
		Frame-EER [%]			Frame-minDCF			$I_{0.7,0.7}$-AUC			$I_{0.7,0.7}$-AUC
without $\mathcal{L}^{scl}$		$\mathbf{x}$	$\mathbf{z}^{MOS}$		$\mathbf{x}$	$\mathbf{z}^{MOS}$		$\mathbf{x}$	$\mathbf{z}^{MOS}$		
	Baseline [3]	15.29	16.61		1.00	1.00		0.30	0.07		0.20
	SUP1	13.34	12.16		**0.93**	**0.93**		0.63	0.43		0.59
	SUP2	11.07	**10.79**		0.95	1.00		0.79	**0.72**		0.76
with $\mathcal{L}^{scl}$		$\mathbf{x}$	$\mathbf{z}^{scl}$	$\tilde{\mathbf{z}}$	$\mathbf{x}$	$\mathbf{z}^{scl}$	$\tilde{\mathbf{z}}$	$\mathbf{x}$	$\mathbf{z}^{scl}$	$\tilde{\mathbf{z}}$	
	CON1	**7.32**	**4.88**	**4.64**	1.00	**0.62**	**0.61**	**0.86**	**0.92**	**0.92**	0.77
	CON2	9.51	7.60	26.7	1.00	0.95	1.00	0.78	0.80	0.05	**0.78**

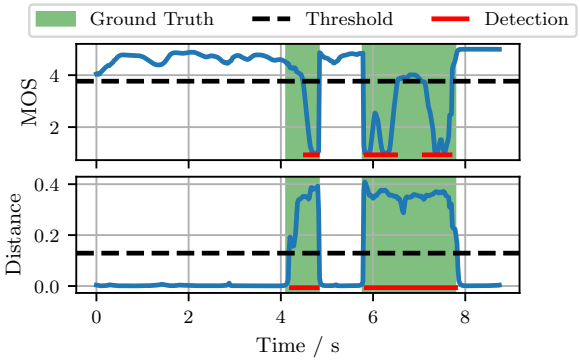

Figure 2: *CON1 MOS-based (upper) and embedding-based (lower) detection results for a single utterance from* `NISQA_TEST_SIM`*-partial-mixup. The threshold is tuned to maximize the intersection-based F1-score on the full test set.*

6.3. Embedding analysis

To analyze whether frame-level embeddings of the same degradation type group together and are distinguishable from embeddings of other degradation types, we use the ground-truth annotations to detect degraded embeddings. We evaluate the embedding purity when using the actual detections from the embedding-based detection in Section 6.4. Given the usually strong correlation of neighboring frames, we reduce the number of frame embeddings to analyze by taking the mean of consecutively detected embeddings.

Table 4 shows the results on `NISQA_TEST_SIM`-partial-mixup when including all 36 degradation combinations. Adding frame-level pseudo-targets to the training does not drastically improve discrimination (Baseline EER: 42.1%, SUP2 EER: 36.3%). Training with $\mathcal{L}_{total}$ significantly improves EER (CON2: 13.5%) and accuracy (SUP2: 7.14%, CON1: 30.3%). Excluding clean frames from the loss gives only marginal improvement in EER (CON1: 15.2%, CON2: 13.5%).

Table 4: *Performance of degradation-type verification and retrieval for all* $K = 36$ *degradation combinations, and using oracle information to detect the degradations. Overall best results for each metric are shown in* ***bold****. Testset:* `NISQA_TEST_SIM`*-partial-mixup.*

Model ID	EER [%]			Accuracy [%]		
	$\mathbf{x}$	$\mathbf{z}^{MOS}$		$\mathbf{x}$	$\mathbf{z}^{MOS}$	
Baseline [3]	42.11	43.86		2.12	0.62	
SUP1	37.81	39.11		0.61	0.62	
SUP2	36.29	44.64		7.14	2.12	
	$\mathbf{z}^{scl}$	$\tilde{\mathbf{z}}$		$\mathbf{z}^{scl}$	$\tilde{\mathbf{z}}$	
CON1	15.17	14.15		**30.30**	27.41	
CON2	**13.56**	14.25		26.79	25.76	

Figure 3 shows how EER and accuracy for CON1 and CON2 behave when constraining the number of simultaneous degradations and thus the number of unique degradation classes. Single degradations ($K = 9$) can be distinguished accurately (CON1 EER: 7.71%, CON2 EER: 0.93%). Here, it becomes evident that excluding clean frames from the set of positive classes during training (CON2) is clearly beneficial for degra-

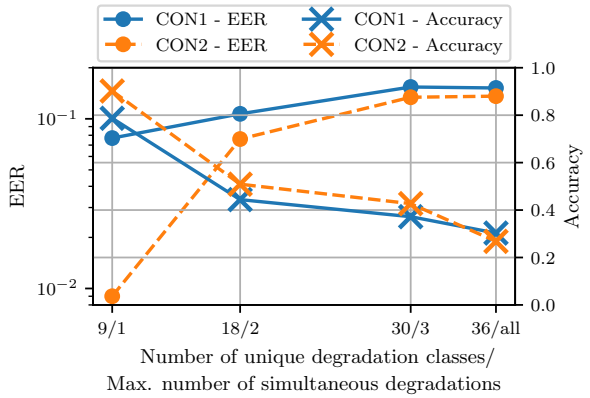

Figure 3: *EER and accuracy when constraining the number of concurrent degradations in* `NISQA_TEST_SIM`*-partial-mixup.*

dation identification. When increasing the maximum number of simultaneous degradations, accuracy starts to drop and EER gets higher, as more classes (two degradations: $K = 18$; three degradations: $K = 30$) share identical distortions, leading to more confusion between classes with shared degradations.

Table 5: *Performance of degradation-type verification and retrieval for unseen degradations using oracle information to detect the degradations. We also report results when restricting to single degradations. Overall best results for each metric and number of simultaneous degradations are shown in* ***bold****. Test-set: LibriAugmented/test-clean-partial-mixup.*

	Single				All			
Model ID	EER [%]		Acc. [%]		EER [%]		Acc. [%]	
	$\mathbf{x}$	$\mathbf{z}^{\text{MOS}}$	$\mathbf{x}$	$\mathbf{z}^{\text{MOS}}$	$\mathbf{x}$	$\mathbf{z}^{\text{MOS}}$	$\mathbf{x}$	$\mathbf{z}^{\text{MOS}}$
Baseline [3]	38.0	38.7	39.7	36.7	36.2	38.6	32.5	27.6
SUP1	33.4	36.5	46.2	43.7	35.2	37.9	34.8	27.4
SUP2	33.6	39.7	46.8	40.3	33.9	38.5	35.2	28.3
	$\mathbf{z}^{\text{scl}}$	$\tilde{\mathbf{z}}$	$\mathbf{z}^{\text{scl}}$	$\tilde{\mathbf{z}}$	$\mathbf{z}^{\text{scl}}$	$\tilde{\mathbf{z}}$	$\mathbf{z}^{\text{scl}}$	$\tilde{\mathbf{z}}$
CON1	19.4	**19.2**	77.5	**78.9**	19.7	**19.5**	65.0	**65.6**
CON2	20.4	20.6	77.2	76.1	19.8	19.9	65.1	64.5

Regarding the results on out-of-domain degradations (Table 5), again, significant improvements can be gained when training with the supervised contrastive loss (SUP2 accuracy: 46.8%/35.2%, CON1 accuracy: 78.9%/65.6%). While accuracy behaves similarly to the in-domain case, the EER is much higher at around 19%, indicating that unseen degradations are harder to separate from one another.

6.4. Joint degradation detection and assignment

So far, the tasks of degradation detection and degradation-type grouping have been treated separately. However, for a downstream application, both are required. Thus, the latent embedding structures of the two contrastive models, CON1 and CON2, are compared. Table 6 shows the comparison with and without oracle detection. Here, additionally, the ARI while ignoring the clean speech class is evaluated (ARI^{dist}), and the accuracy of the clean speech cluster (ACC) is shown. On the in-domain data, for an oracle segmentation, the model trained without the clean speech positive class (CON2) proves significantly better, as also seen in Section 6.3. However, when using

Table 6: *Clustering results in terms of adjusted rand index (ARI) and* ARI^{dist} *(without the clean speech class). ACC denotes the accuracy for the clean speech class for clustering.*

Model ID	Detec.	NISQA			LibriAugmented		
		ARI^{dist}	ARI	ACC	ARI^{dist}	ARI	ACC
CON1	orc.	0.60	–	–	**0.27**	–	–
CON2	orc.	**0.73**	–	–	**0.27**	–	–
CON1	CON1	0.41	**0.68**	**0.94**	**0.27**	0.47	**0.94**
CON2	CON2	0.58	0.13	0.43	0.25	0.34	0.67
CON2	CON1	**0.74**	0.26	0.59	**0.27**	**0.75**	**0.93**

a non-oracle detection, this model significantly degrades. When comparing the ARI with and without scoring the clean speech cluster, this difference is due to the model's inability to consistently output embeddings for clean speech (ARI of 0.13 versus ARI^{dist} of 0.58). In contrast, for model CON1, the clean-speech embeddings are accurately grouped into a single cluster. Using the model CON1 for detection while extracting embeddings with model CON2 mitigates this behavior, but the mismatch between clustering performance persists, so that only half of the clean-speech embeddings are correctly clustered.

For the out-of-domain data, overall clustering accuracy drops. The lower ARI indicates that the model cannot fully separate different degradation types, consistent with the high EERs in Table 5. However, model CON1 still allows accurate discrimination between false positives (clean speech) and degraded embeddings (clean cluster accuracy: 94%). We hypothesize that the drop in clustering performance is due to an inconsistent embedding behavior for degradations unseen during training, but leave a more in-depth analysis for future work. For stability, the inclusion of clean-speech embeddings during training is mandatory to prevent the model from degrading under non-ideal detection. Then, the model yields well-defined clusters for both seen degradations and clean speech. For unseen degradations, this separation is less pronounced, but the embeddings are clearly separate from clean-speech embeddings.

7. Conclusion

We have shown that the detection of local degradation in speech signals can be significantly improved using (i) frame-level pseudo-targets via a partial mix-up data augmentation, (ii) adding a supervised contrastive loss exploiting knowledge about the degradation types, and (iii) switching from MOS-based to embedding-based detection, resulting in near-perfect detection performance on in- and out-of-domain data. The contrastive loss also improved discrimination between degradation types and retrieval of the same types. Further analysis is needed to improve degradation-type discrimination on in- and out-of-domain data and to assess robustness under varying levels of degradation [23]. To this end, an analysis of the similarity among degradation types would be highly beneficial for creating more representative training and test data. Finally, this system can be evolved along the same lines as frame-level audio-language models [11, 12] and trained on textual descriptions to enable automatic assignment of frame embeddings or clusters to degradation types.

8. Acknowledgements

Computational resources were provided by the Paderborn Center for Parallel Computing.

9. References

[1] W.-C. Huang, E. Cooper, and T. Toda, "SHEET: A Multi-purpose Open-source Speech Human Evaluation Estimation Toolkit," in *Proceedings of ISCA Interspeech*, 2025, pp. 2355–2359.

[2] M. Kuhlmann, F. Seebauer, P. Wagner, and R. Haeb-Umbach, "Towards Frame-level Quality Predictions of Synthetic Speech ," in *Proceedings of ISCA Interspeech*, 2025, pp. 2300–2304.

[3] M. Kuhlmann, A. Werning, T. von Neumann, and R. Haeb-Umbach, "Speech quality-based localization of low-quality speech and text-to-speech synthesis artefacts," *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

[4] S. wei Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-Net: An End-to-End Non-intrusive Speech Quality Assessment Model Based on BLSTM," in *Proceedings of ISCA Interspeech*, 2018.

[5] S.-W. Fu, C.-F. Liao, and Y. Tsao, *icassp*, pp. 26–30, 2020.

[6] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proceedings of the International Conference on Machine Learning*, 2019, pp. 2031–2041.

[7] F. Cumlin, X. Liang, V. Ungureanu, C. K. Reddy, C. Schüldt, and S. Chatterjee, "Impairments are clustered in latents of deep neural network-based speech quality models," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[8] N. Turpault, R. Serizel, A. Shah, and J. Salamon, "Sound Event Detection in Domestic Environments with Weakly Labeled Data and Soundscape Synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2019.

[9] J. Hao, S. Ye, C. Lu, F. Dong, and J. Liu, "DCASE 2022 TASK4 CHALLENGE TECHNICAL REPORT," *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2022.

[10] M. Chen, Y. Jin, J. Shao, Y. Liu, B. Peng, and J. Chen, "DCASE 2023 Challenge Task4 Technical Report," *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2023.

[11] Y. Wu, C. Tsirigotis, K. Chen, C.-Z. A. Huang, A. Courville, O. Nieto, P. Seetharaman, and J. Salamon, "FLAM: Frame-wise language-audio modeling," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2025. [Online]. Available: https://openreview.net/forum?id=7fQohcFrxG

[12] P. Primus, F. Schmid, and G. Widmer, "Tacos: Temporally-aligned audio captions for language-audio pretraining," in *Proceedings of Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2025, pp. 1–5.

[13] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6493–6497.

[14] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2001, pp. 749–752.

[15] M. Chinen, F. S. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "Visqol v3: An open source production ready objective speech and audio metric," in *International Conference on Quality of Multimedia Experience (QoMEX)*, 2020, pp. 1–6.

[16] A. Ragano, J. Skoglund, and A. Hines, "NOMAD: Unsupervised Learning of Perceptual Embeddings For Speech Enhancement and Non-Matching Reference Audio Quality Assessment," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1011–1015.

[17] A. Hines and N. Harte, "Speech intelligibility prediction using a neurogram similarity index measure," *Speech Communication*, vol. 54, no. 2, pp. 306–320, 2012.

[18] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 18 661–18 673, 2020.

[19] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization Ability of MOS Prediction Networks," *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8442–8446, 2022.

[20] 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 *Proceedings of ISCA Interspeech*, 2021, pp. 2127–2131.

[21] Takaaki Saeki and Detai Xin and Wataru Nakata and Tomoki Koriyama and Shinnosuke Takamichi and Hiroshi Saruwatari, "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," in *Proceedings of ISCA Interspeech*, 2022, pp. 4521–4525.

[22] Y. Leng, X. Tan, S. Zhao, F. K. Soong, X.-Y. Li, and T. Qin, "MBNET: MOS Prediction for Synthesized Speech with Mean-Bias Network," *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 391–395, 2021.

[23] V. Deng, C. Wang, G. Richard, and B. McFee, "Investigating the sensitivity of pre-trained audio embeddings to common effects," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[24] Y. Tian, L. Fan, P. Isola, H. Chang, and D. Krishnan, "Stablerep: Synthetic images from text-to-image models make strong visual representation learners," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 48 382–48 402, 2023.

[25] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[26] J. S. Chung, J. Huh, S. Mun, M. Lee, H.-S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," in *Proceedings of ISCA Interspeech*, 2020, pp. 2977–2981.

[27] E. Cooper and J. Yamagishi, "How do voices from past speech synthesis challenges compare today?" in *11th ISCA Speech Synthesis Workshop*, 2021, pp. 183–188.

[28] Z.-H. Tan, N. Dehak *et al.*, "rVAD: An unsupervised segment-based robust voice activity detection method," *Computer Speech & Language*, pp. 1–21, 2020.

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[30] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proceedings of Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015, pp. 504–511.

[31] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training," in *Proceedings of ISCA Interspeech*, 2021, pp. 721–725.

[32] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 4214–4217.

[33] M. Przybocki and A. F. Martin, "NIST speaker recognition evaluation chronicles," in *Speaker and Language Recognition Workshop (Odyssey)*, 2004, pp. 15–22.

[34] Çağdaş Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulovic, "A Framework for the Robust Evaluation of Sound Event Detection," *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 61–65, 2020.

[35] S. Cornell, J. Ebbers, C. Douwes, I. Martín-Morató, M. Harju, A. Mesaros, and R. Serizel, "DCASE 2024 task 4: Sound event detection with heterogeneous data and missing labels," *arXiv preprint arXiv:2406.08056*, 2024.

[36] J. Ebbers, R. Haeb-Umbach, and R. Serizel, "Post-Processing Independent Evaluation of Sound Event Detection Systems," in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2023.