International Journal of Speech Technology
https://doi.org/10.1007/s10772-024-10140-6

Comparison of wav2vec 2.0 models on three speech processing tasks

Marie Kunešová[1] · Zbyněk Zajíc[1] · Luboš Šmídl[1] · Martin Karafiát[2]

Received: 6 August 2024 / Accepted: 1 September 2024

Abstract

The current state-of-the-art for various speech processing problems is a sequence-to-sequence model based on a self-attention mechanism known as transformer. The widely used wav2vec 2.0 is a self-supervised transformer model pre-trained on large amounts of unlabeled speech and then fine-tuned for a specific task. The data used for training and fine-tuning, along with the size of the transformer model, play a crucial role in both of these training steps. The most commonly used wav2vec 2.0 models are trained on relatively "clean" data from sources such as the LibriSpeech dataset, but we can expect there to be a benefit in using more realistic data gathered from a variety of acoustic conditions. However, it is not entirely clear how big the difference would be. Investigating this is the main goal of our article. To this end, we utilize wav2vec 2.0 models in three fundamental speech processing tasks: speaker change detection, voice activity detection, and overlapped speech detection, and test them on four real conversation datasets. We compare four wav2vec 2.0 models with different sizes and different data used for pre-training, and we fine-tune them either on in-domain data from the same dataset or on artificial training data created from the LibriSpeech corpus. Our results suggest that richer data that are more similar to the task domain bring better performance than a larger model.

Keywords Speaker change detection · Voice activity detection · Overlapped speech detection · Wav2vec 2.0

1 Introduction

In the past few years, several subfields of machine learning, particularly those related to speech and language, have been overtaken by transformer models. The transformer neural network architecture uses the attention mechanism (Vaswani et al., 2017) and has recently seen great success on a variety of tasks, including but not limited to speech processing (Liu et al., 2021).

For speech processing, one of the currently most successful approaches is wav2vec 2.0 (hereafter referred to as "wav2vec2"), proposed by Baevski et al. (2020). It is a self-supervised transformer model that is pre-trained on large amounts of unlabeled speech data and which can then be fine-tuned for a specific task. Wav2vec2 models have been used for a large variety of different speech processing tasks, such as automatic speech recognition (Baevski et al., 2020; Lehečka et al., 2022), speaker recognition (Vaessen & Van Leeuwen, 2022), and many others (Yang et al., 2021).

In this article, we focus on the use of wav2vec2 models in three basic speech processing tasks that are used in a variety of speech applications: speaker change detection (SCD), voice activity detection (VAD), and overlapped speech detection (OSD). SCD aims to find the points in a conversation where the active speaker changes, while the goal of OSD is to detect intervals where more than one speaker is active at the same time. Voice activity detection simply distinguishes between speech and non-speech. SCD and OSD are both particularly important to speaker diarization (Bullock et al., 2020), automatic speech recognition (Wu et al., 2023), as well as other tasks related to

✉ Marie Kunešová
mkunes@ntis.zcu.cz

✉ Zbyněk Zajíc
zzajic@ntis.zcu.cz

Luboš Šmídl
smidl@kky.zcu.cz

Martin Karafiát
karafiat@fit.vutbr.cz

[1] New Technologies for the Information Society and Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czechia

[2] Faculty of Information Technology, Brno University of Technology, Brno, Czechia

Published online: 10 October 2024

processing multi-speaker audio (Aronowitz & Zhu, 2020). Meanwhile, voice activity detection has a use in nearly all speech processing.

In our earlier paper (Kunešová & Zajíc, 2023), we proposed an approach to speaker change detection, voice activity detection, and overlapped speech detection based on wav2vec2. In another recent publication (Zajíc & Kunešová, 2023), we also explored the capabilities of different pre-trained wav2vec2 models of various sizes and pre-training data by comparing their performance on the SCD task. The article presented here is an extended version of the latter work. While the original conference paper focused only on SCD, here we also add comparisons on the other two tasks: VAD and OSD. We also provide a more detailed analysis of the results, including statistical significance. The performance of the models is evaluated on four widely used corpora of conversational speech.

1.1 Related work

Older approaches for the SCD task include computing a distance between two sliding windows (Rouvier et al., 2013) or detecting differences in pitch (Hogg et al., 2019). More recent works have employed deep learning, using precomputed features based on i-vectors or x-vectors (Aronowitz & Zhu, 2020), Mel-frequency cepstral coefficients (Hogg et al., 2019), spectrograms (Hrúz & Zajíc, 2017), or combinations of multiple types of features (Su et al., 2022). Some authors have even included lexical information gained from automated transcripts (Anidjar et al., 2021; Zajíc et al., 2018) or word embeddings (Jung et al., 2023) for speaker change detection. Various neural network architectures have been applied, such as CNN (Hrúz & Zajíc, 2017), LSTM (Yin et al., 2017), or sequence-level modeling methods (Fan et al., 2022).

The research of OSD has also undergone a similar shift from conventional techniques to deep learning. Where the oldest works relied on hand-crafted features such as LPC residual energy (Boakye et al., 2008), more recent studies use approaches such as convolutional neural networks (Kunešová et al., 2019) or LSTMs (Bullock et al., 2020). The input of the networks can be in the form of MFCCs (Cornell et al., 2020), spectrograms (Kazimirova & Belyaev, 2018), x-vectors (Mateju et al., 2022), or raw waveforms (Bredin & Laurent, 2021; Mariotte et al., 2024).

VAD has been extensively researched for a long time (Ramirez et al., 2007; Tong et al., 2016), but nowadays, it is rarely the main focus by itself. Instead, it usually appears as one part or a by-product of a more complex system, such as in combination with OSD (Cornell et al., 2020; Mariotte et al., 2024) or as part of a speaker diarization system (Han et al., 2021).

2 Pre-trained models and data

Since the original introduction of the wav2vec 2.0 approach, many different pre-trained models have been published, with different numbers of parameters and trained on different datasets.

From this range of options, we chose the following pre-trained models for our evaluation: the original two wav2vec 2.0 models `wav2vec2-base` and `wav2vec2-large`, which are trained on English language data Baevski et al. (2020), the large cross-lingual (XLSR) model `wav2vec2-large-xlsr-53` (Conneau et al., 2021), and finally, to show the efficiency of models trained on different than clean data, also the Czech model `wav2vec2-base-cs-80k-ClTRUS`, which is trained on Czech-language data from a greater variety of different domains, such as radio, telephone, TV shows, and more (Lehečka et al., 2022). The parameters of each model are summarized in Table 1. In subsequent text, we will refer to these models simply as "base", "large", "xlsr-53" and "ClTRUS".

Note that although the pre-training data of the "ClTRUS" model do not match the English language of the datasets used in this article (described below), this should not be an issue—all three examined speech processing tasks are language-independent.

Table 1 Pre-trained wav2vec 2.0 models used in this article

Model	#Tr.	#Par.	Datasets	Hours	Lang.
wav2vec2-base[1]	12	~ 95M	Librispeech	960	English
wav2vec2-large[2]	24	~ 317M	Librispeech	960	English
wav2vec2-large-xlsr-53[3]	24	~ 317M	MLS, CV, BABEL	~ 56k	53 lang.
wav2vec2-base-cs-80k-ClTRUS[4]	12	~ 95M	various	~ 80k	Czech

"#Tr" = the number of transformer blocks, "#Par." = the number of parameters

[1] https://huggingface.co/facebook/wav2vec2-base

[2] https://huggingface.co/facebook/wav2vec2-large

[3] https://huggingface.co/facebook/wav2vec2-large-xlsr-53

[4] https://huggingface.co/fav-kky/wav2vec2-base-cs-80k-ClTRUS

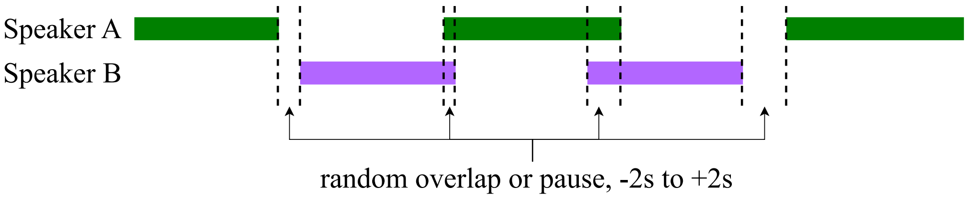

Fig. 1 Artificial training data for SCD. Each colored rectangle represents one single-speaker utterance from the LibriSpeech corpus

To evaluate the effectiveness of the different wav2vec2 models, we tested our system on several widely used English-language conversational speech corpora, which have annotated speaker turns and for which there are recent results from other authors.

Section 2.1 provides a description of the four datasets and their division into training, development, and test data. To study the impact of out-of-domain data, we also used artificially created data as an alternative for training—this is described in Sect. 2.2.

2.1 Real conversation data

For fine-tuning and evaluating our models, we used four English-language corpora of conversational speech:

- The **AMI Meetings Corpus** (AMI) (Carletta, 2007) contains ~100 hours of meetings between 3–5 speakers. We used the "headset mix" recordings for our experiments and followed the official "Full-corpus partition of meetings"[1] for our training set. Evaluation on this corpus was done using the `pyannote` library plugin `pyannote.db.odessa.ami`[2] with its built-in reference annotations (the development and test sets of which are subsets of the "Full-corpus partition's").
- The **CALLHOME American English Speech** (CALLHOME, or "CH") corpus (Canavan et al., 1997) is an 18-hour dataset of narrow-band telephone conversations (converted to 16 kHz for our experiments), typically between two speakers. We used the same training and test sets as Hrúz (Hrúz & Hlaváč, 2018; Hrúz & Zajíc, 2017): 43 and 77 files, respectively. Several files in the training set contain more than two speakers, but the test set consists only of two-speaker conversations. Since the division does not include a development set, and the training set is already fairly small, we evaluate the corpus in this article using a two-fold cross-validation process.
- The **First DIHARD Challenge** (DIHARD I, or "DH-I") dataset (Ryant et al., 2018; Bergelson, 2016) includes audio from 12 challenging domains that range from relatively clean (audiobooks, radio interviews) to noisy recordings (child language recordings, restaurant conversations). As the corpus doesn't have a training set, we split it from the original DIHARD I development data using the partitioning published by Fan et al. (2022)[3].
- The **Second DIHARD Challenge** (DIHARD II, or "DH-II") dataset (Ryant et al., 2019; Bergelson, 2016) is a successor to DIHARD I. We focus on the single-channel track, which contains data from 13 audio sources (some of which only appear in the development set or only in the evaluation set). Most of the contents are identical to DIHARD I, but there are some additions and corrections, including completely recreated annotations for two of the domains. We followed the example of Bullock et al. (2020) and split the development data into 2/3 for training and 1/3 for development (though likely not in the same way).

These four corpora are frequently used for speaker diarization and similar speech-processing tasks, so they are very suitable for testing speaker change detection, which we consider the main focus of our three tasks. They are perhaps slightly less ideal for overlapped speech detection, as the ground truth labels are not always perfectly accurate in this regard, but there are very few (if any) real conversational datasets where that is true.

2.2 Artificial training data

To also compare the effectiveness of the individual wav2vec2 models on out-of-domain data, we used synthetic training data, originally designed in Kunešová et al. (2019) and Kunešová and Zajíc (2023) specifically for the OSD and SCD tasks, respectively. Both were made from the "train-other-500" subset of the LibriSpeech corpus (Panayotov et al., 2015) by artificially combining single-speaker utterances. This way, we can control the speaker change points and overlaps and ensure that reference labels are accurate.

Figure 1 illustrates the creation of the artificial SCD dataset in Kunešová and Zajíc (2023). We generated 500 artificial audio files by always concatenating five different

[1] https://groups.inf.ed.ac.uk/ami/corpus/datasets.shtml

[2] https://github.com/pyannote/pyannote-db-odessa-ami

[3] https://github.com/zhiyunfan/SEQ-SCD/tree/master/data/dihard1

utterances (individual audio files, typically around 5–15 seconds long) from two different speakers in an A-B-A-B-A pattern, with random pauses or overlaps of up to 2 seconds between them. The leading and trailing silence at the start and end of each utterance was linearly tapered to avoid discernible seams. The total duration of this generated dataset was approximately 8 hours.

The creation process of the artificial OSD dataset was somewhat more complicated, so we refer the reader to the original description in Kunešová et al. (2019). This dataset had a total duration of 35 hours.

For VAD, there is no separate artificial dataset—for this task, we simply used the same data as for SCD. The reference VAD labels, like the OSD labels in Kunešová et al. (2019), were previously obtained from the original single-speaker LibriSpeech utterances with a different VAD detector.

3 Methods

In this section, we describe our approach to the three speech-processing tasks: SCD, VAD, and OSD. Unlike our multitask system in Kunešová and Zajíc (2023), here we use a separate model for each task so that they can be evaluated independently, without influencing each other. The structure of the models for each task is identical—the only differences are in the creation of the training labels and in how the outputs are processed.

As described in the above-mentioned paper, we treat each of the three problems as an audio frame classification task, using an approach that was initially proposed in (Kunešová and Řezáčková, 2022) for the detection of prosodic phrases but that can easily be applied to several other speech processing tasks, including SCD, OSD, and VAD.

We use a wav2vec2 model to get a contextual representation of the input signal (raw 16 kHz audio waveform), with an additional last decision layer in the form of one neuron with a linear activation function. This neuron outputs the relevant prediction about each audio frame—i.e., every 20 ms, as per the pre-trained wav2vec2 model. Due to the character of the labeling functions (described in Sects. 3.1 and 3.2), the model is trained for regression, with mean square error loss, rather than a simple binary classification.

The implementation of the model and the fine-tuning process are done using the HuggingFace Transformers library (Wolf et al., 2020), as in our previous work[4]. The model architecture is illustrated in Fig. 2.

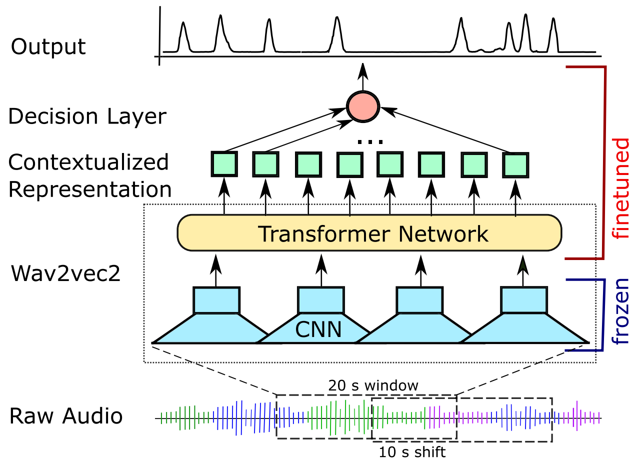

Fig. 2 The basic structure of the wav2vec2 models used for all three tasks. The model outputs a label for each audio frame (every 20 ms)

Because of the high memory requirements of the wav2vec2 models, the input 16 kHz audio signal is given in segments of 20 seconds, with a 10-second overlap between neighboring segments. When the resulting predictions are joined back together for evaluation, we use the middle part of each segment and discard the duplicate 5 s intervals at the edges. This ensures that the prediction for every audio frame is made with sufficient context on both sides.

3.1 Speaker change detection

For the purposes of this work, we define "speaker change" as a point in the audio signal where a currently active speaker stops speaking or a new speaker begins to speak. In other words, we do not only seek boundaries between two different speakers but also between speech and non-speech or between one speaker and multiple.

Reference labels for the SCD task are based on each dataset's annotation files in the Rich Transcription Time Marked (RTTM) format, which is the standard annotation format for speaker diarization. Each line in an RTTM file specifies the time interval and speaker ID of one unbroken speaker turn. In our work, we consider the beginnings and ends of all these intervals as speaker change points, with one minor adjustment: in the training labels used for fine-tuning the model, if two turns of the same speaker have only a small gap (less than one second) between them, we merge the two turns, ignoring the gap. This helps to prevent the model from becoming too sensitive and reporting "speaker changes" even in brief pauses between words. For evaluation, we use the original speaker turns.

In order to deal with time inaccuracies in the human-annotated references, we also use a fuzzy labeling strategy, which we first developed in (Hrúz & Zajíc, 2017): speaker change points are given a reference label with a value of 1,

[4] Our code is available at https://github.com/mkunes/w2v2_audioFrameClassification.

which linearly decreases to zero over an interval of ±0.2 s around each boundary. Audio frames more than 0.2 seconds away from the nearest speaker change point are labeled as 0. An example of this triangular labeling can be seen in Fig. 3.

During the evaluation, we detect speaker change points by first finding peaks (local maxima) in the predicted labels and then applying a threshold—peaks above the threshold are considered speaker change points. However, we also set a minimum distance between detected peaks—if there are multiple peaks within a 0.25 s interval, only the highest one is kept. This brings a very minor but consistent improvement in the results. No other post-processing of the model outputs is performed.

3.2 Voice activity detection and overlapped speech detection

Voice activity detection and overlapped speech detection use an identical approach. Similarly to SCD, the training labels for VAD and OSD use a fuzzy labeling function with values between 0 and 1: 1 indicates speech/overlap, 0 indicates non-speech/non-overlap, and there is a linear slope (width 0.4 s) at the boundaries between the two, as in Kunešová et al. (2019). An example of this labeling can again be seen in Fig. 3.

The information about speech and overlaps is again based on the speaker turns listed in each dataset's annotation. However, unlike SCD, where we remove short pauses within the speech of the same speaker, here we found it better to keep the speaker turns unchanged.

During evaluation, the classifications for each audio frame are obtained with a simple threshold—predicted values higher than the threshold are classified as speech/overlap. We do not perform any post-processing on these two tasks.

4 Results

In this section, we present the experimental results. For each task, we fine-tuned the four pre-trained wav2vec2 models on the training sets of each of the real conversation corpora (separately) and on the artificial data.

The number of epochs was the same for all models—five epochs, as in our previous publication on this topic (Zajíc & Kunešová, 2023). However, while previously we used the same decision thresholds for all models and datasets (selected on the AMI development set), here we tuned the thresholds for each dataset and model separately:

For each fine-tuned model, we found the optimal threshold on the development set based on the lowest or highest value of a specific evaluation metric (see Sect. 4.1). The possible range of the thresholds was -0.1 to 1.1, with a step of 0.01.

For CALLHOME, where, as previously mentioned, we do not have a separate development set, we use a slightly different process, using two-fold cross-validation: We find an optimal threshold on one half of the test set and use it to evaluate the other half and vice versa. The evaluation metrics obtained on each half are then averaged. For this reason, our tables with results always list two thresholds for CALLHOME.

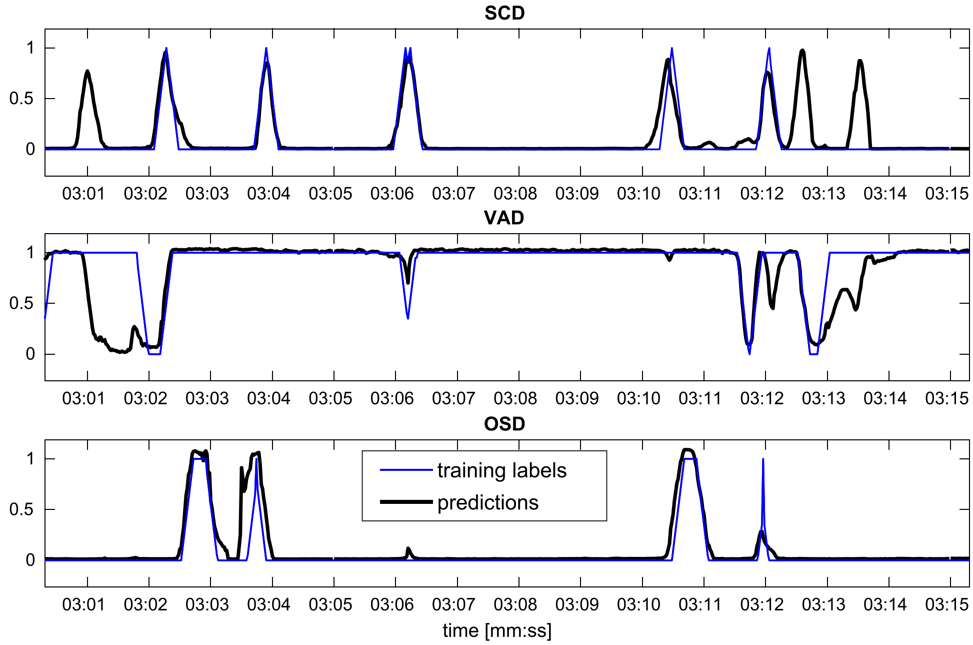

Fig. 3 Predictions and fuzzy labeling functions for all three tasks on the CALLHOME file "en_4431.wav" (model "ClTRUS")

4.1 Evaluation metrics

The predictions for each task were evaluated using the Python library `pyannote.metrics` (Bredin, 2017). For each task, we report evaluation metrics that are commonly used by other authors on the same task.

Speaker change detection was evaluated in terms of audio segmentation, as segment purity (Pur), coverage (Cov), and their harmonic mean (Hn). Purity measures how homogeneous the segments are, and coverage expresses whether each speaker turn is fully contained within one segment. *Hn* is analogous to the F1-score (which typically refers to the harmonic mean of precision and recall), only calculated from different metrics. Purity and coverage are obtained as

$$\text{purity}(S, R) = \frac{\sum_k \max_j |s_k \cap r_j|}{\sum_k |s_k|} \tag{1}$$

$$\text{coverage}(S, R) = \frac{\sum_j \max_k |s_k \cap r_j|}{\sum_j |r_j|}, \tag{2}$$

where $S = \{s_1, \ldots, s_K\}$ is the set of segments (i.e., intervals between speaker changes) found by the system, $R = \{r_1, \ldots, r_J\}$ corresponds to the ground-truth segments, $|r_j|$ is the duration of segment r_j, and $s_k \cap r_j$ denotes the intersection of segments s_k and r_j.

When searching for the optimal threshold on the development set, we chose the setting with the highest *Hn*.

Voice activity detection was evaluated in terms of detection error rate (Err), which is the sum of the missed speech rate (miss) and the false alarm rate (FA). *Miss* is the ratio of speech that was incorrectly labeled as non-speech, while *FA* describes the amount of non-speech that was incorrectly labeled as speech. Both are calculated relative to the total duration of speech in the data. During evaluation, the thresholds for each model were optimized for the lowest *Err* on the development set.

For **overlapped speech detection**, the main metrics were precision, recall, and F1-score. Additionally, we also report accuracy and the detection error rate. The latter is analogous to the detection error rate of VAD—it is calculated relative to the total duration of overlapped speech, and so it can be higher than 100%. For selecting the threshold on the development set, we searched for the highest F1-score.

4.2 Speaker change detection results

The results of speaker change detection for individual corpora can be seen in Table 2. Unlike the preceding conference paper (Zajíc & Kunešová, 2023), where all models used the same threshold of 0.35, here we selected the thresholds for each model and dataset separately. However, the model checkpoints are the same as previously.

As seen in the table, the SCD results on all corpora are very consistent, and with the exception of AMI, the differences between models trained on in-domain or artificial data are relatively small. The consistency of our tested models is also evident from the coverage vs. purity graphs shown in Fig. 4 for all four corpora.

Table 2 Our results (%) for the SCD task with models fine-tuned either on in-domain data or on an artificial dataset

Dataset	Model	In-domain training data				Artificial training data			
		Thresh.	Cov	Pur	Hn	Thresh.	Cov	Pur	Hn
AMI	Base	0.47	92.07	89.25	90.64	0.34	83.36	81.42	*82.38*
	Large	0.32	91.30	90.59	90.94	0.40	81.44	81.95	81.69
	xlsr-53	0.31	91.59	90.83	**91.21**	0.43	85.15	82.36	**83.74**
	ClTRUS	0.27	91.42	90.88	*91.15*	0.42	85.90	78.67	82.13
DH-I	Base	0.39	94.16	89.33	91.68	0.33	92.82	86.17	89.37
	Large	0.28	93.91	90.26	*92.05*	0.52	92.48	86.55	*89.42*
	xlsr-53	0.35	95.56	88.99	**92.16**	0.43	90.36	89.14	**89.75**
	ClTRUS	0.29	94.03	89.80	91.87	0.13	89.85	88.73	89.29
DH-II	Base	0.47	94.08	91.30	92.67	0.50	95.91	85.29	90.29
	Large	0.35	94.75	91.04	92.86	0.60	95.89	85.23	90.25
	xlsr-53	0.30	95.00	91.93	**93.44**	0.50	94.10	88.16	**91.03**
	ClTRUS	0.46	95.87	90.26	*92.98*	0.46	95.83	86.12	*90.72*
CH	Base	0.50/0.40	93.92	92.09	92.99	0.36/0.14	91.80	87.07	89.33
	Large	0.42/0.45	93.53	92.48	93.00	0.30/0.28	90.70	88.02	89.33
	xlsr-53	0.39/0.34	93.64	93.20	*93.42*	0.32/0.23	92.86	89.18	**90.96**
	ClTRUS	0.34/0.31	94.40	92.53	**93.45**	0.11/0.09	92.68	87.69	*90.10*

The best result for each dataset is shown in bold text; the second best in italics

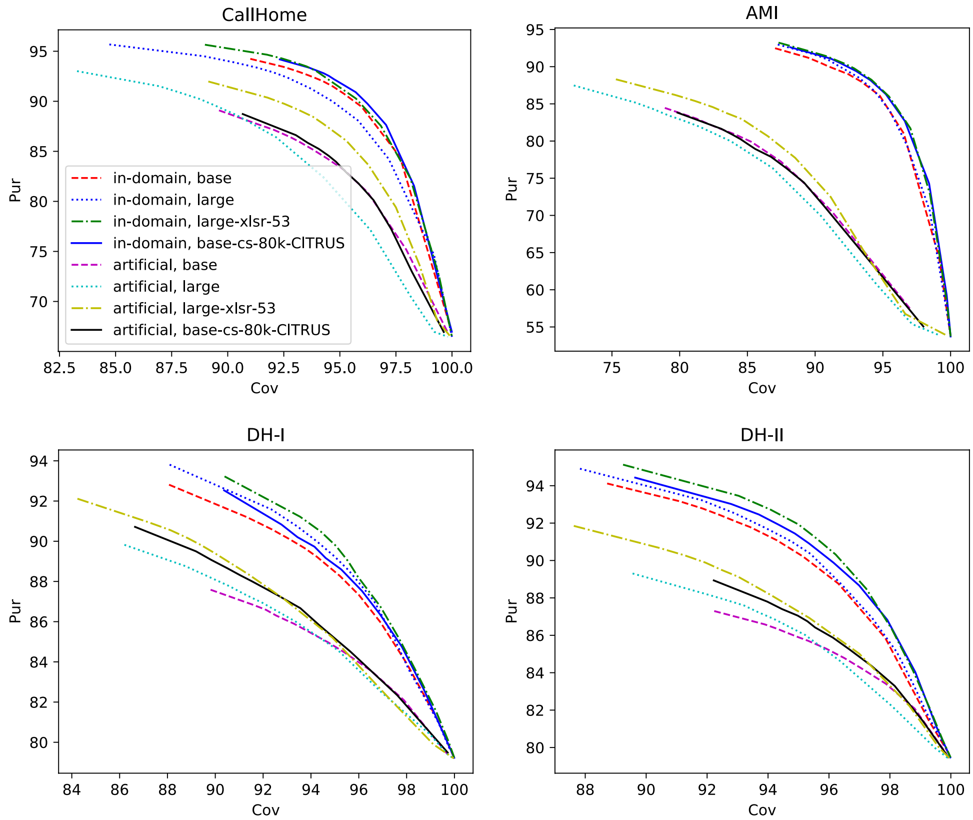

Fig. 4 Coverage vs. Purity on the SCD task for different thresholds with models fine-tuned on in-domain or artificial data. (Zajíc & Kunešová, 2023)

4.3 Voice activity detection results

Table 3 shows the results of voice activity detection. Here, the differences between datasets and between in-domain vs. artificial data are larger.

The results on AMI again stand out: when models are trained on in-domain data, AMI has the lowest error rate of all four datasets. But with models trained on artificial data, voice activity detection on AMI completely failed - the "optimal" threshold is the lowest one, when *everything* is classified as "speech".

Upon closer examination, we found that the VAD models trained on artificial data are overly sensitive—even some pauses between words are classified as non-speech. This increased sensitivity applies to some degree to all four corpora but appears to be the most pronounced on AMI. However, we feel that this is, in essence, not so much an error as a mismatch in the level of granularity between the predictions and the reference annotations. It likely stems from the nature of the VAD labels used for the artificial data, which were themselves obtained from the original single-speaker utterances with a different voice activity detector. The models trained on such labels would naturally reflect the sensitivity of the earlier VAD.

On the other hand, the models trained on artificial data also classify some non-speech sounds, such as breath and laughter, as speech, even though they are not marked as such in the ground-truth annotations. Such sounds are particularly common in the AMI dataset but not in the artificial data. The models trained on AMI do not have this issue. This will also be discussed in Sect. 5.1.

4.4 Overlapped speech detection results

Finally, the results of overlapped speech detection are shown in Table 4.

As seen in the table, overlap detection on the DIHARD I and DIHARD II datasets performed poorly. However, this is in line with previous OSD results on these very challenging data (Miasato Filho et al., 2018; Bullock et al., 2020).

The results with models trained on artificial training data were also substantially worse than when trained on in-domain data, especially for the AMI corpus. This suggests that the artificial dataset for OSD needs to be improved so that it better represents real conversations.

For AMI in particular, we suspect that our artificial OSD data do not model the overlapped speech in AMI very well—for example, each artificial file combines speech from two speakers, but AMI also contains overlaps between three or more individuals. Additionally, there are some instances of overlapping laughter, which is not marked as speech but may be detected as such during both VAD and OSD.

Table 3 Our results (%) for the VAD task for different models fine-tuned for 5 epochs

Dataset	Model	Thresh.	Err	Miss	FA	Acc
(a) Models fine-tuned on in-domain data						
AMI	Base	0.29	5.54	2.13	3.41	95.48
	Large	0.46	5.09	2.22	2.87	95.85
	xlsr-53	0.34	*4.98*	1.88	3.10	95.94
	ClTRUS	0.45	**4.89**	1.97	2.93	96.01
DH-I	Base	0.15	13.16	4.71	8.45	90.03
	Large	0.10	13.57	4.93	8.64	89.72
	xlsr-53	0.16	**11.67**	5.65	6.02	91.16
	ClTRUS	0.07	*11.94*	3.85	8.09	90.96
DH-II	Base	0.07	12.54	3.91	8.63	90.64
	Large	0.10	12.77	4.90	7.87	90.47
	xlsr-53	0.08	**11.64**	4.62	7.02	91.32
	ClTRUS	0.09	*12.05*	4.29	7.76	91.01
CH	Base	0.42/0.49	8.51	4.02	4.48	92.61
	Large	0.37/0.45	*8.19*	4.19	4.00	92.88
	xlsr-53	0.44/0.40	**8.18**	4.65	3.52	92.89
	ClTRUS	0.45/0.50	8.34	3.79	4.55	92.75
(b) Models fine-tuned on artificial data						
AMI	Base	– 0.10	*22.71*	0.00	22.70	81.50
	Large	– 0.03	**22.69**	0.00	22.69	81.51
	xlsr-53	– 0.10	22.73	0.03	22.71	81.47
	ClTRUS	– 0.10	*22.71*	0.00	22.71	81.50
DH-I	Base	0.07	**18.10**	7.75	10.35	86.29
	Large	0.04	18.42	9.39	9.03	86.04
	xlsr-53	0.06	*18.23*	8.46	9.77	86.19
	ClTRUS	0.09	19.34	8.87	10.47	85.34
DH-II	Base	0.05	18.25	5.76	12.49	86.38
	Large	– 0.00	**17.57**	4.56	13.01	86.89
	xlsr-53	0.04	*17.85*	6.60	11.26	86.68
	ClTRUS	0.06	18.77	6.59	12.18	86.00
CH	Base	0.03/0.04	11.42	7.29	4.13	90.07
	Large	– 0.01/0.00	**10.38**	5.30	5.08	90.98
	xlsr-53	0.02/0.02	*10.90*	6.15	4.75	90.53
	ClTRUS	0.04/0.05	11.31	5.86	5.45	90.17

The best result for each dataset is shown in bold text; the second best in italics

5 Analysis and discussion

5.1 Performance on different datasets

Before we compare the different pre-trained wav2vec2 models, it would be a good idea to also discuss the differences in performance between the four datasets used for evaluation.

As observed in the previous sections, the AMI corpus has the greatest difference between models trained on AMI data and those trained on artificial data. This difference may stem from the character of the AMI dataset—it contains many instances of non-speech sounds, such as breath and laughter, which are not present in the artificial data. Additionally, as previously mentioned, these sounds are typically not marked as speech in the ground-truth annotations but may be detected as such by our models. This would naturally lead to increased errors, especially in the VAD and OSD predictions. For models that were trained on the AMI training data, the issue is mitigated as they were able to adapt to the AMI annotations and learn what should and should not be considered speech.

By contrast, CALLHOME has the most consistent results. This might be partially helped by the fact that, unlike the other datasets, our CALLHOME test set always has only two speakers in each conversation—just like the artificial data.

Finally, the DIHARD I and DIHARD II datasets had reasonable results on SCD and VAD but fared very poorly

Table 4 Our results (%) for the OSD task for different models fine-tuned for 5 epochs

Dataset	Model	Thresh.	Prec	Rec	F1	Acc	Err
(a) Models fine-tuned on in-domain data							
AMI	Base	0.19	76.11	81.62	78.77	93.83	44.00
	Large	0.14	78.19	82.07	80.08	94.28	40.83
	xlsr-53	0.14	78.47	82.75	**80.55**	94.40	39.96
	ClTRUS	0.15	77.98	83.04	*80.43*	94.34	40.40
DH-I	Base	0.13	42.05	50.10	45.73	91.98	118.94
	Large	0.12	42.78	53.17	47.41	92.05	117.95
	xlsr-53	0.10	48.68	49.33	**49.00**	93.08	102.68
	ClTRUS	0.11	45.18	51.96	*48.33*	92.51	111.10
DH-II	Base	0.13	49.93	61.08	54.95	93.38	100.17
	Large	0.32	51.58	60.72	55.78	93.64	96.27
	xlsr-53	0.17	54.00	58.89	*56.34*	93.97	91.28
	ClTRUS	0.16	50.91	65.15	**57.16**	93.55	97.66
CH	Base	0.16/0.14	64.77	68.50	66.58	94.62	68.77
	Large	0.19/0.20	70.42	69.21	*69.79*	95.30	59.96
	xlsr-53	0.19/0.19	70.78	69.11	**69.92**	95.35	59.48
	ClTRUS	0.16/0.16	68.06	69.27	68.64	95.04	63.33
(b) Models fine-tuned on artificial data							
AMI	Base	0.03	60.32	60.81	60.56	88.90	79.20
	Large	0.01	65.53	62.51	63.98	90.14	70.37
	xlsr-53	0.00	60.51	76.23	**67.47**	89.70	73.52
	ClTRUS	0.04	76.86	56.28	*64.98*	91.50	60.66
DH-I	Base	0.06	38.62	48.25	42.90	91.34	128.43
	Large	0.02	34.94	57.00	43.32	89.95	149.16
	xlsr-53	0.01	38.25	60.34	**46.82**	90.76	137.05
	ClTRUS	0.05	40.99	52.91	*46.19*	91.69	123.28
DH-II	Base	0.07	43.81	51.27	47.24	92.44	114.49
	Large	0.07	42.97	54.85	48.19	92.21	117.95
	xlsr-53	0.03	46.45	59.27	**52.08**	92.79	109.07
	ClTRUS	0.06	47.20	55.50	*51.02*	92.96	106.57
CH	Base	0.17/0.17	56.16	61.14	58.53	93.23	86.64
	Large	0.19/0.19	55.91	65.08	60.14	93.28	86.25
	xlsr-53	0.11/0.10	66.74	68.46	**67.59**	94.87	65.67
	ClTRUS	0.12/0.10	65.07	64.78	*64.90*	94.55	70.07

The best result for each dataset is shown in bold text; the second best in italics

on the OSD task—although we are not the only ones to face this issue (Miasato Filho et al., 2018; Bullock et al., 2020). This could be influenced by several factors: Firstly, the audio files do not come from a single source. Rather, as mentioned in section 2.1, these are collections of data from several different speech corpora with different characteristics, some of which only appear in the development or test sets. This makes it more difficult to find a single model setting or threshold that would work well on the entire dataset, as opposed to tuning them on each subset separately (Zajíc et al., 2018). A part of the problem might also be in the reference labels—the DIHARD I annotations, in particular, are known to be imperfect and had been partially replaced in DIHARD II (Ryant et al., 2019). This may also account for the substantial difference between our OSD results on DIHARD I and DIHARD II.

For a comparison with other systems from different state-of-the-art articles, we also present Table 5, showing the best results on the selected corpora we could find in the literature. Here, we can see that the SCD results of all our models surpass the previous works, while OSD and VAD are at least comparable.

5.2 Statistical analysis of differences between pre-trained models

The tables with our results on the three tasks show that model “xlsr-53” nearly always achieved the best results of

Table 5 Previously reported results (%) for the three tasks on different corpora

Task	Dataset	Publication	Cov	Pur	Hn
SCD	AMI	Su et al. (2022)	91.75	85.68	88.61
SCD	DH-I	Fan et al. (2022)	92.56	86.24	89.29
SCD	DH-II	Bredin et al. (2020)	93.7	86.8	–
SCD	CH	Hrúz and Hlaváč (2018)	72.57	72.57	–
Task	Dataset	Publication	Miss	FA	Err
VAD	AMI	Bredin and Laurent (2021)	3.2	3.6	6.8
VAD	DH-I	Miasato Filho et al. (2018)	9.3	–	12.7
VAD	DH-II	Bredin et al. (2020)	4.2	5.7	9.9
VAD	CH	Landini et al. (2022)	3.23	4.03	–
Task	Dataset	Publication	Prec	Rec	F1
OSD	AMI	Chen et al. (2024)	–	–	81.76
OSD	DH-I	Miasato Filho et al. (2018)	62.1	9.9	17.1
OSD	DH-II	Chen et al. (2024)	–	–	61.37
OSD	CH	Chen et al. (2024)	–	–	70.13

For AMI and CH, the test sets may have been different from ours

the four pre-trained models, while "base" tended to be the worst.

However, to determine whether the differences between the pre-trained models are meaningful, we calculated the statistical significance of our results using the Wilcoxon signed-rank test. This was done in the following manner:

For each task, we calculated the F1-scores (OSD), Hn values (SCD), or detection error rates (VAD) for each individual file in the test sets of the AMI corpus (22 files), CALLHOME (77 files), and DIHARD II (194 files from 11 separate audio domains). We did not include DIHARD I in this comparison, as most of its audio files are also part of DIHARD II.

Then, for every task and every combination of two models A and B, we performed the Wilcoxon signed-rank test comparing the two sets of per-file results, with the alternative hypothesis that model A is better than model B, i.e., that its F1-scores or Hn values are larger (OSD, SCD) or that its detection error rates are lower (VAD).

The resulting p-values (without any correction for multiple comparisons) can be found in Table 6.

From this and the previous tables, we can safely conclude that models "ClTRUS", "large" and "xlsr-53" all surpass the performance of the base wav2vec2 model. Additionally, "ClTRUS" appears to be better than "large".

Model "xlsr-53" also achieved better results than "large" in most cases and usually scored best of all four models (see Tables 2, 3, and 4), but according to Table 6, the difference is statistically significant only for the SCD task.

For models "large" and "xlsr-53", the fact that they are better than "base" is an expected result—they have three times as many parameters, and "xlsr-53" was additionally pre-trained on a much larger set of training data (see Table 1).

Table 6 Statistical significance of our results with models trained on in-domain data

	↓ A / B →	Base	Large	xlsr-53	ClTRUS
SCD	Base	–	0.828166	1	1
	Large	0.171834	–	1	0.999993
	xlsr-53	**3.057e-17**	**5.544e-21**	–	**0.000373**
	ClTRUS	**3.878e-07**	**7.103e-06**	0.999627	–
VAD	Base	–	0.921325	0.950494	0.999039
	Large	0.078675	–	0.586875	0.999607
	xlsr-53	**0.049506**	0.413125	–	0.949927
	ClTRUS	**0.000961**	**0.000393**	0.050073	–
OSD	Base	–	0.999997	0.999996	1
	Large	**2.634e-06**	–	0.705546	0.982099
	xlsr-53	**4.097e-06**	0.294454	–	0.798838
	ClTRUS	**4.494e-09**	**0.017901**	0.201162	–

Obtained using the Wilcoxon signed-rank test (without correction for multiple comparisons), with the alternative hypothesis that model A is better than model B. Values of $p < 0.05$ are shown in bold

The results for the "ClTRUS" model are more interesting. The model has the same size and architecture as the "base" model and differs only in the data used for pre-training, yet it also consistently brings better results.

The "base" and "large" models were trained mainly on clean LibriSpeech data and had not been exposed to real wild acoustic conditions such as those in the tested datasets. On the other hand, the "ClTRUS" model saw many different kinds of "wild" data during the pre-training phase, and the fine-tuning on in-domain data can benefit from this.

Similarly, the larger “xlsr-53” model, which was pre-trained on more variable data from a few different datasets, also supports this trend.

6 Conclusion

Our main goal in this article was to determine the effect of using models pre-trained on data from a variety of real acoustic conditions as opposed to models pre-trained on relatively clean data. To this end, we compared the performance of four different wav2vec2 models with an additional decision layer when fine-tuned for the SCD, VAD, and OSD tasks.

A comparison of these models on four datasets of conversational speech shows us the importance of in-domain data not only in the fine-tuning phase but also in the self-supervised pre-training phase. Based on the results, we believe that using richer data for pre-training the models brings a bigger gain than using larger models with more parameters.

Wav2vec2 is a relatively complex model with a high computation cost, so using it for some of these tasks—especially the relatively simple task as voice activity detection—could be considered an overkill. However, we wish to use this approach in a transcription system in combination with existing ASR (Lehečka et al., 2022), where the first wav2vec2 layers can be shared.

Acknowledgements This research was supported by the Czech Ministry of Interior, Project ROZKAZ (VJ01010108). Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Funding Open access publishing supported by the National Technical Library in Prague.

Data Availability The data used in this article were obtained from the following datasets: AMI Meeting Corpus (https://groups.inf.ed.ac.uk/ami/corpus/index.shtml), LibriSpeech (https://www.openslr.org/12), CALLHOME American English Speech (LDC Catalog No. LDC97S42), First DIHARD Challenge (LDC Catalog No. LDC2019S09, LDC2019S10, LDC2019S13, LDC2019S12), and Second DIHARD Challenge (LDC Catalog No. LDC2021S10, LDC2021S11, LDC2022S06, LDC2022S07).

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Anidjar, O. H., Lapidot, I., Hajaj, C., Dvir, A., & Gilad, I. (2021). Hybrid speech and text analysis methods for speaker change detection. *IEEE/ACM Transactions on Audio Speech and Language Processing, 29*, 2324–2338. https://doi.org/10.1109/TASLP.2021.3093817

Aronowitz, H., & Zhu, W. (2020). Context and uncertainty modeling for online speaker change detection. In *Proceedings of the international conference on acoustics, speech, and signal processing* (*ICASSP*) (pp 8379–8383). https://doi.org/10.1109/ICASSP40776.2020.9053280

Baevski, A., Zhou, Y., Mohamed, A., Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449–12460. https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

Bergelson, E. (2016). Bergelson seedlings HomeBank Corpus. https://doi.org/10.21415/T5PK6D

Boakye, K., Trueba-Hornero, B., Vinyals, O., & Friedland, G. (2008). Overlapped speech detection for improved speaker diarization in multiparty meetings. In *2008 IEEE international conference on acoustics, speech and signal processing* (pp. 4353–4356). https://doi.org/10.1109/ICASSP.2008.4518619

Bredin, H. (2017). Pyannote.metrics: A toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems. In *Proceedings of the Interspeech* (pp. 3587–3591). https://doi.org/10.21437/INTERSPEECH.2017-411

Bredin, H., & Laurent, A. (2021). End-to-end speaker segmentation for overlap-aware resegmentation. In *Proceedings of the Interspeech* (pp. 3111–3115). https://doi.org/10.21437/Interspeech.2021-560

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., & Gill, M. P. (2020). Pyannote.audio: Neural building blocks for speaker diarization. In *Proceedings of the international conference on acoustics, speech, and signal processing* (*ICASSP*) (pp. 7124–7128). https://doi.org/10.1109/ICASSP40776.2020.9052974

Bullock, L., Bredin, H., & Garcia-Perera, L. P. (2020). Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection. In *Proceedings of the international conference on acoustics, speech, and signal processing* (*ICASSP*) (pp. 7114–7118). https://doi.org/10.1109/ICASSP40776.2020.9053096

Canavan, A., Graff, D., & Zipperlen, G. (1997). CALLHOME American English Speech, LDC97S42. https://doi.org/10.35111/exq3-x930

Carletta, J. (2007). Unleashing the killer corpus: Experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation, 41*(2), 181–190. https://doi.org/10.1007/S10579-007-9040-X

Chen, Z., Han, B., Wang, S., Qian, Y. (2024). Attention-based encoder-decoder end-to-end neural diarization with embedding enhancer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32*, 1636–1649. https://doi.org/10.1109/TASLP.2024.3366756

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2021). Unsupervised cross-lingual representation learning for speech recognition. In *Proceedings of the Interspeech* (pp. 2426–2430). https://doi.org/10.21437/Interspeech.2021-329

Cornell, S., Omologo, M., Squartini, S., & Vincent, E. (2020). Detecting and counting overlapping speakers in distant speech scenarios. In *Proceedings of the Interspeech* (pp. 3107–3111). https://doi.org/10.21437/Interspeech.2020-2671

Fan, Z., Dong, L., Cai, M., Ma, Z., & Xu, B. (2022). Sequence-level speaker change detection with difference-based continuous integrate-and-fire. *IEEE Signal Processing Letters, 29*, 1551–1554. https://doi.org/10.1109/LSP.2022.3185955

Han, E., Lee, C., & Stolcke, A. (2021). BW-EDA-EEND: Streaming END-TO-END neural speaker diarization for a variable number of speakers. In *Proceedings of the international conference on acoustics, speech, and signal processing* (*ICASSP*) (pp. 7193–7197). https://doi.org/10.1109/ICASSP39728.2021.9414371

Hogg, A. O. T., Evers, C., & Naylor, P. A. (2019). Speaker change detection using fundamental frequency with application to multi-talker segmentation. In *Proceedings of the international conference on acoustics, speech, and signal processing* (*ICASSP*) (pp. 5826–5830). https://doi.org/10.1109/ICASSP.2019.8682924

Hrúz, M., & Hlaváč, M. (2018). LSTM neural network for speaker change detection in telephone conversations. *Speech and Computer SPECOM 2018 Lecture Notes in Computer Science, 11096*, 226–233. https://doi.org/10.1007/978-3-319-99579-3_24

Hrúz, M., & Zajíc, Z. (2017). Convolutional neural network for speaker change detection in telephone speaker diarization system. In *Proceedings of the international conference on acoustics, speech, and signal processing* (*ICASSP*) (pp. 4945–4949). https://doi.org/10.1109/ICASSP.2017.7953097

Jung, J. W., Seo, S., Heo, H. S., Kim, G., Kim, Y. J., Kwon, Y. K., Lee, M., & Lee, B. J. (2023). Encoder-decoder multimodal speaker change detection. In *Proceedings of the Interspeech* (pp. 5311–5315). https://doi.org/10.21437/Interspeech.2023-2289

Kazimirova, E., & Belyaev, A. (2018). Automatic detection of multi-speaker fragments with high time resolution. In *Proceedings of the Interspeech* (pp. 1388–1392). https://doi.org/10.21437/Interspeech.2018-1878

Kunešová, M., & Řezáčková, M. (2022). Detection of prosodic boundaries in speech using wav2vec 2.0. *Text, Speech, and Dialogue TSD 2022 Lecture Notes in Computer Science, 13502*, 377–388. https://doi.org/10.1007/978-3-031-16270-1_31

Kunešová, M., & Zajíc, Z. (2023). Multitask detection of speaker changes, overlapping speech and voice activity using wav2vec 2.0. In *Proceedings of the international conference on acoustics, speech, and signal processing* (*ICASSP*) (pp. 1–5). https://doi.org/10.1109/ICASSP49357.2023.10094972

Kunešová, M., Hrúz, M., Zajíc, Z., & Radová, V. (2019). Detection of overlapping speech for the purposes of speaker diarization. In *Speech and computer SPECOM 2019 lecture notes in computer science* (Vol. 11658, pp. 247–257). https://doi.org/10.1007/978-3-030-26061-3_26

Landini, F., Lozano-Diez, A., Diez, M., & Burget, L. (2022). From simulated mixtures to simulated conversations as training data for end-to-end neural diarization. In *Proceedings of the Interspeech, 2022* (pp. 5095–5099). https://doi.org/10.21437/Interspeech.2022-10451

Lehečka, J., Švec, J., Pražák, A., & Psutka, J. V. (2022). Exploring capabilities of monolingual audio transformers using large datasets in automatic speech recognition of Czech. In *Proceedings of the Interspeech* (pp. 1831–1835). https://doi.org/10.21437/INTERSPEECH.2022-10439

Liu, A. T., Li, S. W. W., & Lee, Hy Y. (2021). TERA: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Transactions on Audio Speech and Language Processing* 29:2351–2366. https://doi.org/10.1109/TASLP.2021.3095662

Mariotte, T., Larcher, A., Montrésor S, & Thomas, J. H. (2024). Channel-combination algorithms for robust distant voice activity and overlapped speech detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32*, 1859–1872. https://doi.org/10.1109/TASLP.2024.3369531

Mateju, L., Kynych, F., Cerva, P., Malek, J., & Zdánský, J. (2022). Overlapped speech detection in broadcast streams using x-vectors. In *Proceedings of the Interspeech* (pp. 4606–4610). https://doi.org/10.21437/Interspeech.2022-81

Miasato Filho, V. A., Silva, D. A., & Depra Cuozzo, L. G. (2018). Joint discriminative embedding learning, speech activity and overlap detection for the DIHARD speaker diarization challenge. In *Proceedings of the Interspeech, 2018* (pp. 2818–2822). https://doi.org/10.21437/Interspeech.2018-2304

Panayotov, V., Chen, G., Povey, D., Khudanpur, S. (2015). LibriSpeech: An ASR corpus based on public domain audio books. In *Proceedings of the international conference on acoustics, speech, and signal processing* (*ICASSP*) (pp. 5206–5210). https://doi.org/10.1109/ICASSP.2015.7178964

Ramirez, J., Górriz, J. M., & Segura, J. C. (2007). Voice activity detection. Fundamentals and speech recognition system robustness. In *Robust speech recognition and understanding* (pp. 1–22). https://doi.org/10.5772/4740

Rouvier, M., Dupuy, G., Gay, P., Khoury, E., Merlin, T., & Meignier, S. (2013). An open-source state-of-the-art toolbox for broadcast news diarization. In *Proceedings of the Interspeech* (pp. 1477–1481). https://doi.org/10.21437/INTERSPEECH.2013-383

Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., & Liberman, M. (2018). First DIHARD challenge evaluation plan. Tech. rep., Linguistic Data Consortium, https://catalog.ldc.upenn.edu/docs/LDC2019S09/first_dihard_eval_plan_v1.3.pdf

Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., & Liberman, M. (2019). The second DIHARD diarization challenge: Dataset, task, and baselines. In *Proceedings of interspeech* (pp. 978–982). https://doi.org/10.21437/Interspeech.2019-1268

Su, H., Zhao, D., Dang, L., Li, M., Wu, X., Liu, X., & Meng, H. (2022). A multitask learning framework for speaker change detection with content information from unsupervised speech decomposition. In *Proceedings of the international conference on acoustics, speech, and signal processing* (*ICASSP*) (pp. 8087–8091). https://doi.org/10.1109/ICASSP43922.2022.9746116

Tong, S., Gu, H., & Yu, K. (2016). A comparative study of robustness of deep learning approaches for VAD. In *Proceedings of the international conference on acoustics, speech, and signal processing* (*ICASSP*) (pp. 5695–5699). https://doi.org/10.1109/ICASSP.2016.7472768

Vaessen, N., & Van Leeuwen, D. A. (2022). Fine-tuning wav2vec2 for speaker recognition. In *2022 IEEE international conference on acoustics, speech and signal processing* (*ICASSP 2022*) (pp. 7967–7971). https://doi.org/10.1109/ICASSP43922.2022.9746952

Vaswani, A. (2017). Attention is all you need. In *Proceedings of the 31st international conference on neural information processing systems* (*NIPS'17*) (pp. 5998–6008). https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Davison, J. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45). https://doi.org/10.18653/v1/2020.emnlp-demos.6

Wu, J., Chen, Z., Hu, M., Xiao, X., & Li, J. (2023). Speaker change detection for transformer transducer ASR. In *Proceedings of the international conference on acoustics, speech, and signal* *processing (ICASSP)* (pp. 1–5). https://doi.org/10.1109/ICASSP49357.2023.10096361, arXiv: 2302.08549

Yang, S. W., Chi, P. H., Chuang, Y. S., Lai, C. I. J., Lakhotia, K., Lin, Y. Y., Liu, A. T., Shi, J., Chang, X., Lin, G. T., & Huang, T. H. (2021). SUPERB: Speech processing Universal PERformance Benchmark. In *Proceedings of the Interspeech* (pp. 1194–1198). https://doi.org/10.21437/Interspeech.2021-1775

Yin, R., Bredin, H., & Barras, C. (2017). Speaker change detection in broadcast TV using bidirectional long short-term memory networks. In *Proceedings of the Interspeech 2017* (pp. 3827–3831). https://doi.org/10.21437/Interspeech.2017-65

Zajíc, Z., & Kunešová, M. (2023). Comparison of wav2vec 2.0 transformer models for speaker change detection. In M. Abbas, & A. A. Freihat (Eds.), *Proceedings of the 6th international conference on natural language and speech processing* (*ICNLSP 2023*) (pp. 233–238). https://aclanthology.org/2023.icnlsp-1.23

Zajíc, Z., Kunešová, M., Zelinka, J., & Hrúz, M. (2018). ZCU-NTIS speaker diarization system for the DIHARD 2018 challenge. In *Proceedings of the Interspeech* (pp. 2788–2792). https://doi.org/10.21437/Interspeech.2018-1252

Zajíc, Z., Soutner, D., Hrúz, M., Müller, L., & Radová, V. (2018). Recurrent neural network based speaker change detection from text transcription applied in telephone speaker diarization system. *Text, Speech, and Dialogue TSD 2018 Lecture Notes in Computer Science, 11107*, 342–350. https://doi.org/10.1007/978-3-030-00794-2_37

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.