# Comparison of Wav2vec 2.0 Transformer Models for Speaker Change Detection

**Zbyněk Zajíc** and **Marie Kunešová**
New Technologies for the Information Society and Department of Cybernetics,
Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic
zzajic@ntis.zcu.cz, mkunes@ntis.zcu.cz

## Abstract

The state-of-the-art for various speech tasks is a sequence-to-sequence model based on a self-attention mechanism known as Transformer. The broadly used Wav2vec 2.0 is a self-supervised transformer model pre-trained on large unlabeled datasets and subsequently fine-tuned for a particular task. The data, along with the size of the transformer model, play a crucial role in both these training steps. In this paper, we utilize Wav2vec 2.0 models for finding the speaker change in a speech signal. Our goal is to compare different model sizes with different training datasets to show that data similar to the task domain bring better performance than larger models. The speaker change detection task was tested on four real conversation corpora with consistent top results.

## 1 Introduction

Speaker change detection (SCD) is the task of finding the point in a conversation where the speaker is changing. It is a basic speech-processing task that is relevant to various speech applications such as speaker diarization (Bullock et al., 2020; Kunešová et al., 2017; Zajíc et al., 2016), automatic speech recognition (Wu et al., 2023), and other tasks related to processing multi-speaker audio (Aronowitz and Zhu, 2020; Zajíc et al., 2018).

Legacy approaches for the SCD task include computing a distance between two sliding windows (Rouvier et al., 2013), detecting differences in pitch (Hogg et al., 2019), or using precomputed features based on i/x-vectors (Aronowitz and Zhu, 2020), Mel-frequency cepstral coefficients (MFCCs) (Hogg et al., 2019), spectrograms (Hrúz and Zajíc, 2017), and combinations of multiple types of features (Su et al., 2022), even including lexical information gained from automated transcripts (Anidjar et al., 2021; Zajíc et al., 2018) or word embeddings (weon Jung et al., 2023) for speaker change detection. Different neural network model architectures have been applied, such as LSTM (Hrúz and Hlaváč, 2018), CNN (Hrúz and Zajíc, 2017), or sequence-level modeling methods (Fan et al., 2022). Nowadays, the transformer network concept uses the attention mechanism of deep learning (Vaswani et al., 2017), which has recently seen great success on a variety of tasks, including but not limited to speech processing (Liu et al., 2021). The main benefit is self-supervised learning on unlabeled data.

In this paper, we investigate the wav2vec 2.0 (Baevski et al., 2020) framework in an end-to-end approach for SCD, first proposed in our previous paper (Kunešová and Zajíc, 2023), where it was shown to achieve state-of-the-art results. The main focus of this paper is to explore the capabilities of different pre-trained wav2vec 2.0 models of various sizes. The results are evaluated on four conversational speech corpora broadly used in the SCD task.

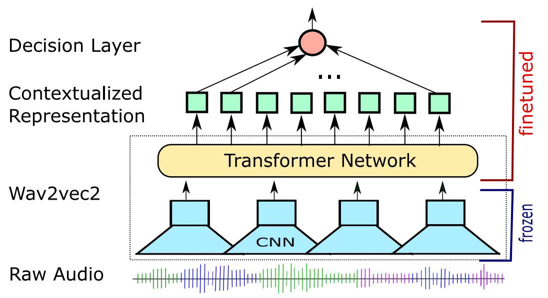


Figure 1: Illustration of the multitask wav2vec 2.0 detector of speaker changes. The model outputs a label for each audio frame (every 20 ms).

## 2 Wav2vec 2.0 models

Self-supervised audio transformers are known to scale well with the size of pre-training data. Wav2vec 2.0 (hereafter referred to as “wav2vec2”) is a transformer-based self-supervised framework for speech representation, which has been used for a wide range of speech processing tasks, such as automatic speech recognition (Lehečka

Table 1: Pre-trained wav2vec2 models used in this paper.

| Model | #Trans. | #Param. | Datasets | Hours | Lang. |
|---|---|---|---|---|---|
| wav2vec2-base (Baevski et al., 2020) | 12 | $\sim 95M$ | Librispeech | 960 | English |
| wav2vec2-large (Baevski et al., 2020) | 24 | $\sim 317M$ | Librispeech | 960 | English |
| wav2vec2-large-xlsr-53 (Conneau et al., 2021) | 24 | $\sim 317M$ | MLS, CV, BABEL | $\sim 56k$ | 53 lang. |
| wav2vec2-base-cs-80k-ClTRUS (Lehečka et al., 2022) | 12 | $\sim 95M$ | various | $\sim 80k$ | Czech |

et al., 2022) and many others (Yang et al., 2021). There is a huge family of these models with different numbers of parameters trained on different datasets. From this zoo, we pick four models[1] for our evaluation: two that were used in (Kunešová and Zajíc, 2023) – the base model `wav2vec2-base` and the large cross-lingual (XLSR) model `wav2vec2-large-xlsr-53`, plus two others. We added the English large model `wav2vec2-large` and, 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 data from a greater variety of different domains (Lehečka et al., 2022). Their parameters are summarized in Table 1.

## 3 Speaker Change Detection (SCD) task

Speaker change in the SCD task is defined as a point in the audio signal where the speaker changes to another speaker, silence, or overlapping speech. The point where a speaker starts to speak after a silence is also a speaker change.

SCD is generally language-independent because language can be seen as one part of the speaker's characteristics. We try to discriminate these speakers from each other (to find their change). On the other hand, the discrepancy in the train and test acoustic domains plays a significant role in the speech representation by the end-to-end model.

The absence of a large quantity of labeled data needed for the deep learning approach forces us to use a self-supervised model as wav2vec2.

### 3.1 SCD model

As described in our previous paper (Kunešová and Zajíc, 2023), we treat the SCD problem as an audio frame classification task. We use the wav2vec2 model to get a contextual representation of the input signal, with an additional last decision layer as a speaker change detector. The outputs from the transformer are fully connected to the decision layer (one neuron with a linear activation function), which outputs information about the speaker changes in each audio frame every 20 ms, as per the pre-trained wav2vec2 model. Due to the character of the labeling function (see Section 3.2), the model is trained for regression (with mean square error loss) rather than a simple binary classification. The AdamW algorithm was used as an optimizer except for the wav2vec2-large model, where an Adamax provided more stable training behavior.

For the fine-tuning on SCD-labeled data, only the first CNN layer is frozen. For this step, we are using the HuggingFace Transformers (Wolf et al., 2020) library, as in our aforementioned previous paper[2]. The system's architecture is in Figure 1.

Because of the high memory requirements of the wav2vec2 models, the 16 kHz input signal is given in segments of 20 seconds, with a 10-second overlap between segments. Then 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 there is always sufficient context on both sides of a potential speaker change point.

### 3.2 Reference labels for SCD

Reference labels for the SCD task are based on the annotation files in the Rich Transcription Time Marked (RTTM) format (i.e., 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: during fine-tuning, 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.

Additionally, in order to deal with time inaccu-

[1] Downloaded from `https://huggingface.co/facebook/wav2vec2-base`, .../wav2vec2-large, .../wav2vec2-large-xlsr-53 and `https://huggingface.co/fav-kky/wav2vec2-base-cs-80k-ClTRUS`

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

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

| **Evaluated corpus** | **Feature model** | **In-domain train data** | | | **Artificial train data** | | |
|---|---|---|---|---|---|---|---|
| | | Cov | Pur | F1 | Cov | Pur | F1 |
| AMI | wav2vec2-base | 90.94 | 90.06 | 90.50 | 83.45 | 81.34 | 82.38 |
| | wav2vec2-large | 91.52 | 90.31 | 90.91 | 80.25 | 82.77 | 81.49 |
| | wav2vec2-large-xlsr-53 | 92.20 | 90.39 | **91.28** | 83.45 | 83.76 | **83.61** |
| | wav2vec2-base-cs-80k-ClTRUS | 92.41 | 89.97 | 91.18 | 85.02 | 79.61 | 82.22 |
| DH-I | wav2vec2-base | 93.74 | 89.65 | 91.65 | 92.93 | 86.09 | 89.38 |
| | wav2vec2-large | 94.98 | 89.25 | 92.03 | 91.29 | 87.32 | 89.26 |
| | wav2vec2-large-xlsr-53 | 95.56 | 89.00 | **92.16** | 89.43 | 89.79 | **89.61** |
| | wav2vec2-base-cs-80k-ClTRUS | 94.61 | 89.17 | 91.81 | 91.04 | 88.31 | 89.65 |
| DH-II | wav2vec2-base | 92.93 | 92.09 | 92.51 | 95.00 | 85.90 | 90.22 |
| | wav2vec2-large | 94.75 | 91.04 | 92.86 | 93.67 | 87.24 | 90.34 |
| | wav2vec2-large-xlsr-53 | 95.59 | 91.19 | **93.33** | 92.46 | 89.51 | **90.96** |
| | wav2vec2-base-cs-80k-ClTRUS | 94.88 | 91.45 | 93.13 | 95.29 | 86.75 | 90.82 |
| CallHome | wav2vec2-base | 93.48 | 92.70 | 93.09 | 92.83 | 86.38 | 89.49 |
| | wav2vec2-large | 92.62 | 93.36 | 92.99 | 89.62 | 89.40 | 89.51 |
| | wav2vec2-large-xlsr-53 | 93.51 | 93.49 | 93.50 | 93.79 | 88.47 | **91.05** |
| | wav2vec2-base-cs-80k-ClTRUS | 94.51 | 92.54 | **93.51** | 94.51 | 84.55 | 89.25 |

racies in the human-annotated references, we also use a fuzzy labeling strategy, which we first developed in (Hrúz and Zajíc, 2017): speaker change points are given a reference label with a value of 1, which linearly decreases to zero over an interval of $\pm 0.2$ s around each boundary. Audio frames more than 0.2 s away from the nearest speaker change point are labeled as 0.

During 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. In this paper, unlike (Kunešová and Zajíc, 2023), we also set a minimum distance between detected peaks as 0.25 s – if there are multiple peaks within 0.25 s, only the highest one is kept (this brings a very minor but consistent improvement in F1-score). However, the fine-tuned "base" and "xlsr-53" models themselves were identical to the previous work. No other post-processing of the model outputs is performed.

## 4 Datasets

To evaluate the effectiveness of different wav2vec2 models, we tested our system on several widely used English-language conversational speech corpora, which have annotated speaker turns for SCD evaluation.

The tested corpora were the following: AMI Meetings Corpus (**AMI**) (Carletta, 2007), the American English subset of the CallHome (**CallHome**) (Canavan et al., 1997), and the First and Second DIHARD Challenge data (**DH-I**) (Ryant et al., 2018; Bergelson, 2016) and (**DH-II**) (Ryant et al., 2019; Bergelson, 2016).

To also compare the effectiveness of the individual wav2vec2 models on out-of-domain data, we designed a synthetic training dataset in (Kunešová et al., 2019; Kunešová and Zajíc, 2023), made from the LibriSpeech corpus. This way, we can control the speaker change points and also ensure that reference labels are accurate.

## 5 Results and discussion

Predicted speaker change points were evaluated in terms of audio segmentation, as segment purity (Pur), coverage (Cov), and F1-score, using the Python library `pyannote.metrics`[3] (Bredin, 2017). Purity measures how homogeneous the segments are, and coverage expresses whether each speaker turn is fully contained within one segment. F1-score is the harmonic mean of the two.

Results[4] for individual corpora can be seen in Table 2. We used identical settings for all our models and corpora. We set these values in such a way as to obtain high F1 scores on the AMI development set across all models that were trained or evaluated on AMI – as five training epochs and a threshold of 0.35. The consistency of our tested models is evident from the Coverage vs. Purity graph in Figure 2 for all four corpora.

[3]Downloaded from: `https://pyannote.github.io/`

[4]Unlike our results in (Kunešová and Zajíc, 2023), a minimum distance between peaks (0.25 $s$) is applied in this study.

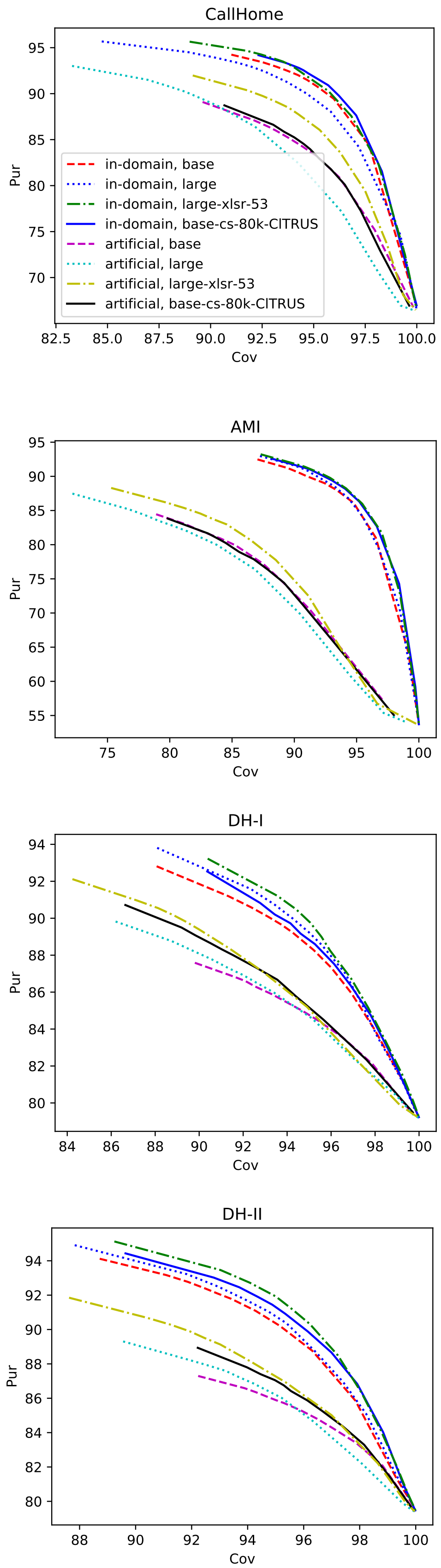


Figure 2: Cov vs. Pur for different thresholds with models fine-tuned on in-domain or artificial data.

Table 3: Previously reported SCD results (%) on different corpora, with models fine-tuned on in-domain data.

| **Corpus and SCD method** | Cov | Pur | F1 |
|---|---|---|---|
| AMI (Su et al., 2022) | 91.75 | 85.68 | 88.61 |
| AMI (Fan et al., 2022) | 89.81 | 83.92 | 86.76 |
| AMI (Bredin et al., 2020) | 84.2 | 90.4 | – |
| DH-I (Fan et al., 2022) | 92.56 | 86.24 | 89.29 |
| DH-II (Bredin et al., 2020) | 93.7 | 86.8 | – |
| CallH. (Hrúz and Hlaváč, 2018) | 72.57 | 72.57 | – |

In comparing the *base* and *large* models, where the number of parameters and the amount of pre-training data are substantially different, the larger models (three times more parameters), especially "xlsr-53", expectedly outperform the base model. The results for the "ClTRUS" model are more interesting. The better-trained "ClTRUS" model with the same architectural size as the base model also consistently brings better results, and is mostly better than the larger models on in-domain data.

The base and large models were trained mainly on clean Librispeech data and are unfamiliar with real wild acoustics conditions in tested data. On the other hand, the "ClTRUS" model saw "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 trained on more variable data from a few different datasets, also supports this trend.

For a comparison with other systems from different state-of-the-art articles, we present Table 3, showing the best results on the selected corpora we could find in the literature.

## 6 Conclusion

In this paper, we tested four different wav2vec2 models with an additional decision layer for the SCD task. Wav2vec2 is a relatively complex model with a high computation cost, but we want 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. The results of our system with all the tested models surpass all previous results on the same datasets. A comparison of these models shows us the importance of in-domain data not only in fine-tuning phase but also in the self-supervised pre-training phase. According to the results, we believe that richer data for pre-training the models brings more gain than bigger models.

## Acknowledgements

This research was supported by the Czech Ministry of Interior, project ROZKAZ (VJ01010108) and by the Czech Ministry of Education, Youth and Sports, Project No. (LM2023062) LINDAT/CLARIAH-CZ. Computational resources provided by the e-INFRA CZ project (ID:90254) were greatly appreciated.

## References

Or Haim Anidjar, Itshak Lapidot, Chen Hajaj, Amit Dvir, and Issachar Gilad. 2021. Hybrid speech and text analysis methods for speaker change detection. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 29:2324–2338.

Hagai Aronowitz and Weizhong Zhu. 2020. Context and uncertainty modeling for online speaker change detection. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 8379–8383.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Elika Bergelson. 2016. Bergelson Seedlings HomeBank Corpus.

Hervé Bredin. 2017. pyannote.metrics: A toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems. In *Proc. Interspeech*, pages 3587–3591.

Hervé Bredin et al. 2020. pyannote.audio: Neural building blocks for speaker diarization. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 7124–7128.

Latané Bullock, Hervé Bredin, and Leibny Paola Garcia-Perera. 2020. Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 7114–7118.

Alexandra Canavan, David Graff, and George Zipperlen. 1997. CALLHOME American English Speech, LDC97S42. In *LDC Catalog*. Linguistic Data Consortium.

Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation*, 41(2):181–190.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Proc. Interspeech*, pages 2426–2430.

Zhiyun Fan, Linhao Dong, Meng Cai, Zejun Ma, and Bo Xu. 2022. Sequence-level speaker change detection with difference-based continuous integrate-and-fire. *IEEE Signal Processing Letters*, 29:1551–1554.

Aidan O.T. Hogg, Christine Evers, and Patrick A. Naylor. 2019. Speaker change detection using fundamental frequency with application to multi-talker segmentation. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5826–5830.

Marek Hrúz and Miroslav Hlaváč. 2018. LSTM neural network for speaker change detection in telephone conversations. *Speech and Computer. SPECOM 2018. Lecture Notes in Computer Science*, 11096:226–233.

Marek Hrúz and Zbyněk Zajíc. 2017. Convolutional neural network for speaker change detection in telephone speaker diarization system. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4945–4949.

Marie Kunešová, Marek Hrúz, Zbyněk Zajíc, and Vlasta Radová. 2019. Detection of overlapping speech for the purposes of speaker diarization. *Speech and Computer. SPECOM 2019. Lecture Notes in Computer Science*, 11658:247–257.

Marie Kunešová and Zbyněk Zajíc. 2023. Multitask detection of speaker changes, overlapping speech and voice activity using wav2vec 2.0. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1–5.

Marie Kunešová, Zbyněk Zajíc, and Vlasta Radová. 2017. Experiments with segmentation in an online speaker diarization system. *Text, Speech, and Dialogue. TSD 2017. Lecture Notes in Computer Science*, 10415:429–437.

Jan Lehečka, Jan Švec, Aleš Pražák, and Josef V. Psutka. 2022. Exploring capabilities of monolingual audio transformers using large datasets in automatic speech recognition of Czech. In *Proc. Interspeech*, pages 1831–1835.

Andy T. Liu, Shang-Wen Wen Li, and Hung-yi Yi Lee. 2021. TERA: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 29:2351–2366.

Mickael Rouvier, Ǵregor Dupuy, Paul Gay, Elie Khoury, Teva Merlin, and Sylvain Meignier. 2013. An open-source state-of-the-art toolbox for broadcast news diarization. In *Proc. Interspeech*, pages 1477–1481.

Neville Ryant et al. 2018. First DIHARD Challenge evaluation plan. Technical report, Linguistic Data Consortium.

Neville Ryant et al. 2019. The Second DIHARD Diarization Challenge: Dataset, task, and baselines. In *Proc. Interspeech*, pages 978–982.

Hang Su et al. 2022. A multitask learning framework for speaker change detection with content information from unsupervised speech decomposition. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 8087–8091.

Ashish Vaswani et al. 2017. Attention is all you need. In *Proc. 31st International Conference on Neural Information Processing Systems (NIPS'17)*, pages 5998–6008.

Jee weon Jung et al. 2023. Encoder-decoder multimodal speaker change detection. In *Proc. Interspeech*, pages 5311–5315.

Thomas Wolf et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Jian Wu, Zhuo Chen, Min Hu, Xiong Xiao, and Jinyu Li. 2023. Speaker change detection for transformer transducer ASR. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1–5.

Shu-wen Yang et al. 2021. SUPERB: Speech processing Universal PERformance Benchmark. In *Proc. Interspeech*, pages 1194–1198.

Zbyněk Zajíc, Marie Kunešová, and Vlasta Radová. 2016. Investigation of segmentation in i-vector based speaker diarization of telephone speech. *Speech and Computer. SPECOM 2016. Lecture Notes in Computer Science*, 9811:411–418.

Zbyněk Zajíc, Daniel Soutner, Marek Hrúz, Luděk Müller, and Vlasta Radová. 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.

Zbyněk Zajíc et al. 2018. First insight into the processing of the language consulting center data. *Speech and Computer. SPECOM 2018. Lecture Notes in Computer Science*, 11096:778–787.