INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

GL-SSD: Global and Local Speech Style Disentanglement by vector quantization for robust sentence boundary detection in speech stream

Kuncai Zhang[1], *Wei Zhou*[1], *Pengcheng Zhu*[1],*Haiqing Chen*[1]

[1]Alibaba Group, China

kuncai.zkc@alibaba-inc.com, fayi.zw@alibaba-inc.com,
tangju.zpc@alibaba-inc.com,haiqing.chenhq@alibaba-inc.com

Abstract

Sentence boundary detection (SBD) in speech, aimed at segmenting the sentence units from the audio speech, plays a significant role in a broad range of tasks such as automatic speech recognition and speech translation. Previous studies have explored the solution based on basic acoustic features and high level semantic representation. Although widely studied, sentence boundary detection still remains a challenge when applied to different speech styles, including the global style and local style. To improve the robustness of SBD in the scene of different speech styles, we propose Global and Local Speech Style Disentanglement (GL-SSD) by vector quantization from the raw speech and incorporate the disentangled style representations into the semantic representation. Relevant experiments demonstrate the superior performance of the proposed method compared to other recent mainstream methods.

Index Terms: sentence boundary detection, Speech and audio segmentation, vector quantization

1. Introduction

Sentence boundary detection (SBD) [1] is a crucial part in the human understanding of spoken language. AS shown in Figure 1, SBD aims to detect the sentence boundaries and segment the proper sentence units from the raw speech. It is an important step for many speech-related tasks, such as automatic speech recognition (ASR) [2] and speech translation (ST) [3].

Methods for SBD can be generally classified into text-based and acoustic-based. The text-based SBD is usually applied in the cascaded speech translation systems [4]. It depends on the transcribed texts or the lexical features to predict the sentence boundary. The acoustic-based SBD usually works for the end-to-end speech translation systems [5], where the transcribed texts are not available and the sentence boundaries need to be detected with the raw audio. In this study, we limit the scope to the acoustic-based SBD, where only the acoustic information is available for boundary detection.

SBD in speech is challenging because the speaker may leave an incomplete sentence, pause for a period of different lengths, or speak for a long time without pausing [6]. The sentence boundary is closely related to the speech styles including speech rate, pause length, pause timing , and so on [7, 8], which are varied with the different speakers. Based on the comprehensive observation of different speeches, we further find that there are global speech styles and local speech styles. The global speech styles are related to the habits or characters of the speaker. While the local speech styles are related to the specific situation throughout the speaking process. Different global and local styles lead to different speech rates and pause lengths, and will further influence the semantic sentence boundaries.

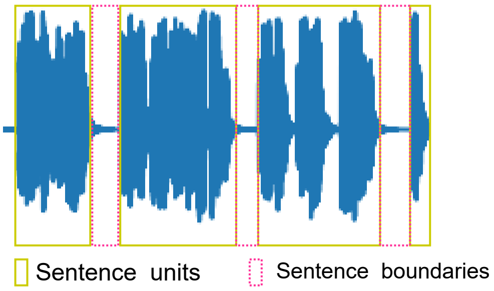

Figure 1: *Introduction of the sentence boundary detection*

In this study, we propose to disentangle the global and local speech styles from the raw speech to improve the robustness of SBD in the scene of different speech styles. The proposed method first adopts a self-supervised approach for Global and Local Speech Style Disentanglement (GL-SSD) by vector quantization from the raw speech. Then the disentangled representations of global and local speech styles are incorporated into the semantic representation to train a sequence classification model for robust sentence boundary detection. We verified the superior performance of the proposed GL-SSD across four different spoken languages of French, Spanish, Portuguese and Italian. Relevant experiments on the Multilingual TEDx (mTEDx) dataset [9] show that the proposed methodology outperforms other mainstream methods by a large margin.

2. Related work

Early studies on SBD considered modeling the basic acoustic features with traditional machine learning methods like hidden Markov model (HMM)[10, 11], conditional random fields (CRF) [12, 13, 14], decision trees (DT) [10] and support vector machines (SVM)[15, 16, 17].

Recently, some researchers attempt to improve the performance of SBD with the help of voice activity detection (VAD) [18] algorithms or wav2vec 2.0 model [19]. VAD-based methods [20, 21] detect the silences in the audio and get the sentence boundaries according to specific rules. Inaguma et al. [20] utilized the heuristic concatenation of VAD segments up to a fixed length to detect the sentence boundary. In [21], Gaido et al. proposed a VAD-hybrid method by giving more importance to the target segments' length than to the detected pauses to improve the performance. Wav2vec 2.0 based methods [22, 23, 24] utilized the semantic representations pretrained from large scale of speech corpus to predict the sentecne boundaries. Marie

10.21437/Interspeech.2023-558

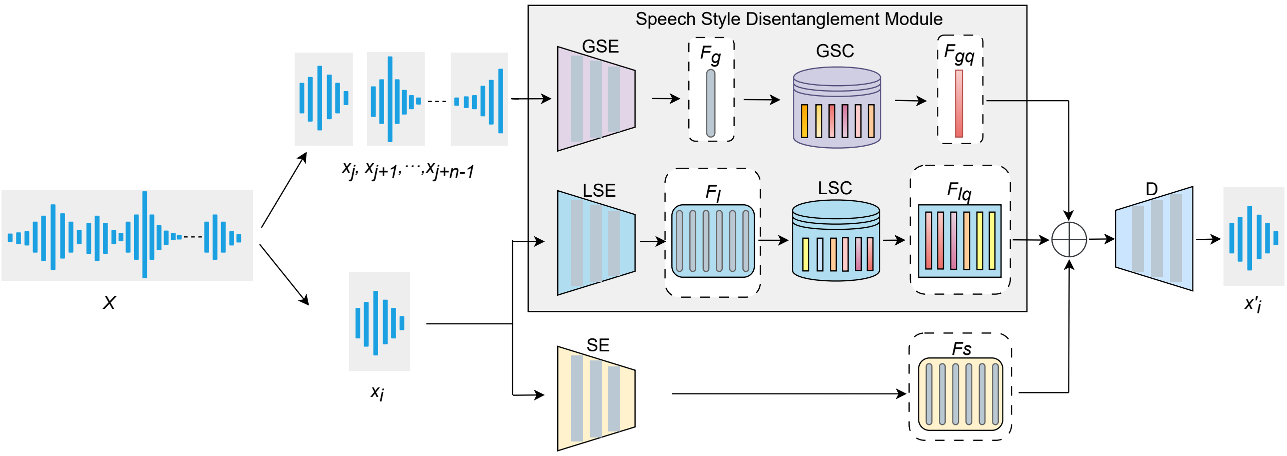
Speech Style Disentanglement Module
GSE
F_g
GSC
F_{gq}
LSE
F_l
LSC
F_{lq}
D
SE
Fs
$x_j, x_{j+1}, \cdots, x_{j+n-1}$
X
x_i
x'_i

Figure 2: *Overall architecture of the proposed GL-SSD. The dotted boxes represent the intermediate outputs of different modules. GSE, LSE and SE represent global style encoder, local style encoder and semantic encoder, respectively. GSC and LSC represent global style codebook and local style codebook, respectively. More details are shown in Section 3.1.1.*

Kunešová et al. [22] propose to utilize the wav2vec 2.0 pretrained model to train a sequence classification model to predict whether each wave frame is speech unit or a boundary. In [23] and [24], the wav2vec 2.0 pretrained model was utilized to segment the sentence units for end-to-end speech translation.

The VAD-based methods often split audio at inappropriate boundaries because they mainly segment the speech boundaries based on long pause which does not coincide with the semantic sentence boundaries. In addition, the hyper-parameter of VAD-based methods is hard to adjust to a proper value which is adaptive to various cases. The methods based on wav2vec 2.0 model usually perform better than the VAD-based methods because they can utilize the semantic representation pretrained from a large scale of speech corpus. The semantic representation can help to predict a better semantic sentence boundary to some degree. However, the methods based on wav2vec 2.0 model still struggle to deal with the case of various speech styles since they could hardly adapt to different speech styles of various speakers. In order to deal with the different speech styles, we propose a robust method for SBD in speech stream by incorporating the disentangled representations of global and local speech styles into the semantic representation.

3. Methodology

The proposed method adopts a two-stage strategy to train a robust sentence boundary detection model. The first stage is to train the disentangled representations of global and local speech styles by vector quantization. The second stage is to train a frame-level sequence classification model for SBD based on the learned hybrid representations.

3.1. Train the disentangled representations of global and local speech styles

3.1.1. Model architecture

The first stage trains the disentangled representation of global and local speech styles with a self-supervised manner. The overall model architecture is shown in Figure 2. The core module of the proposed GL-SSD is the Speech Style Disentanglement Module (SSDM) which disentangles the global and local speech styles by vector quantization. Given a long untrimmed speech of a specific speaker, a speech snippet x_i is randomly sampled from the untrimmed speech as the objective sample to be reconstructed. This speech snippet is input to a semantic encoder SE and a local style encoder LSE, thus produce the semantic representation $F_s \in \mathbb{R}^{d \times t}$ and the local style feature $F_l \in \mathbb{R}^{d \times t}$. Here, t is the length of the feature and d is the dimension size of the feature. Several other n speech snippets (x_j,x_{j+1},...,x_{j+n-1}) are also randomly sampled from the same raw speech of the same speaker. These speech snippets are input to the global style encoder GSE and then averaged to produce the global style feature $F_g \in \mathbb{R}^{d \times 1}$. The global style feature F_g is vector-quantized through the global style codebook GSC, and the local style feature F_l is vector-quantized through the local style codebook LSC. Thus F_g is transformed to the vector-quantized global style representation $F_{gq} \in \mathbb{R}^{d \times 1}$ and F_l is transformed to the vector-quantized local style representation $F_{lq} \in \mathbb{R}^{d \times t}$. Referred to [25], the vector-quantization is a process which can quantize a sequence of continuous data into the closest discrete code. Specifically, given Z as a sequence of continuous data, that is, $Z = z_0, z_1, ..., z_t$. Then the vector-quantization process can be described as:

$$VQ(Z) = q_0, q_1, ..., q_t \tag{1}$$

$$q_j = \underset{q \in codebook}{\arg\min} \ (q - z_j) \tag{2}$$

Then the vector-quantized global style representation F_{gq}, the vector-quantized local style representation F_{lq} are incorporated into the semantic representation F_s by a feature fusion module. We copy the F_{gq} for t times to match with F_s and F_{lq}, then sum them up to get the fused representation. The fused representation is passed to the decoder D to predict the reconstructed speech sample $x_i^{'}$.

We use the front 15 layer of the pretrained wav2vec 2.0 model of the 300m parameter version [1] as the semantic encoder SE to extract the semantic feature. The local style encoder LSE and decoder D are the same as the encoder and decoder from SoundStream [26], respectively. For the implementation of global style encoder GSE, we add a global average pooling layer on top of the SoundStream's [26] encoder.

[1] https://huggingface.co/facebook/wav2vec2-xls-r-300m

3.1.2. Optimized loss function

Similar to [25], the proposed GL-SSD is trained by a self-supervised manner and the optimized loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{vq}^{g} + \mathcal{L}_{vq}^{l} + \alpha\mathcal{L}_{com}^{g} + \beta\mathcal{L}_{com}^{l} \quad (3)$$

where the $\mathcal{L}_{rec}$ is the reconstructed loss, $\mathcal{L}_{vq}^{g}$ and $\mathcal{L}_{vq}^{l}$ are the vector-quantization losses of the global codebook and local codebook, $\mathcal{L}_{com}^{g}$ and $\mathcal{L}_{com}^{l}$ are the commitment losses of the global encoder and local encoder. α and β are the parameter to balance the different loss sources. The reconstructed loss, vector-quantization loss and commitment loss can be derived as:

$$\mathcal{L}_{rec} = \|x_i - x_i^{'}\|_1^1 \quad (4)$$

$$\mathcal{L}_{vq} = \sum_j \|sg[z_j] - q_j\|_2^2 \quad (5)$$

$$\mathcal{L}_{com} = \sum_j \|z_j - sg[q_j]\|_2^2 \quad (6)$$

where sg stands for the stop-gradient operator.

3.2. Train the SBD model

The second stage is to train the SBD model based on the learned hybrid representations of the vector-quantized global style representation and vector-quantized local style representation. Specifically, we replace the decoder D in Figure 2 with a sequence classification module, which takes the fused representations of F_{gq}, F_{lq} and F_s as input, and output the boundary/non-boundary predictions of each frame. The sequence classification module is composed of a transformer layer and a fully-connected classification layer. During training, we fixed the encoders and codebooks, only update the parameters of the sequence classification module. The training is optimized with the cross entropy loss.

4. Experimental setup

4.1. Data

We conduct the experiments on the mTEDx dataset [9]. The mTEDx dataset is an open dataset composed of multilingual corpus of TED talks and the speech styles are various among different speakers. We verified the performance of the proposed method for SBD on four speech subsets across four different spoken languages of French (Fr), Spanish (Es), Portuguese (Pt) and Italian (It). We randomly sampled about 10% data from the training set to expand and diversify the test set because of the extreme imbalance of the original ratio for training and test set. All of the audio data are converted to mono and sampled at a rate of 16 kHz. We label each frame of the audio as sentence boundary/non-boundary according to the start-end time from the annotations.

4.2. Details of training and inference

For the first stage, we train the disentangled representations of global-local speech styles by the self-supervised manner using the training set. For each of the long untrimmed talks, one snippet of about 20 seconds is randomly sampled from the raw speech as the sample to be reconstructed. And about 10 other snippets are sampled from the same speech to extract the global speech style. The dimension size d of the codebook embedding is 16 for both of the global and local codebooks. The size of global codebook is 64 and the size of local codebook is 128. α and β in the loss function are both set to 0.1. The model is trained for about 100 epochs with the Adam optimizer [27] and an initial learning rate of 5×10^{-5}.

After completing the training of the disentangled representations, we add a learnable sequence classification module at the top of the learned hybrid representations to train a sentence boundary detection model. During training, the data input to the encoders is processed in the same way as the first stage. At this stage, we only train the sequence classification module and keep the bottom encoders and codebooks fixed. The model is trained for about 20 epochs with the Adam optimizer and the initial learning rate is 3×10^{-4} which decays with cosine annealing. We choose the best model according to the performance on the valid set and evaluate the performance on the test set.

At the inference time, given an audio waveform of a long untrimmed speech, we predict the confidence of boundary/non-boundary for each frame with a non-overlapping rolling window of a fixed length. The fixed length of the rolling window is 20s, the same as training. Different from the training process, the global style encoder, the local style encoder and the semantic encoder are all input with the same speech snippet which is to be inferred. In order to obtain a more reliable result, we perform the rolling inference with two different offsets and then average the result of each frame.

4.3. Evaluation

In order to verify the superior performance of the proposed method for SBD, we compared the proposed method with the recent mainstream methods including VAD-hybrid and wav2vec 2.0 based methods. The F1 score is computed to evaluate the performance for sentence boundary detection.

Baseline 1: VAD-hybrid. For the VAD-hybrid based method, the experimental setups are simlar to [21]. The widely-used tool WebRTC's VAD [2] is used for silence detection. We tune the frame length parameter in (10, 20, 30) ms and the aggressiveness parameter in (1, 2, 3), where higher values mean more aggressive splits. We set the max length parameter to 30 and tune the min length parameter in range of (0,30).

Baseline 2: Wav2vec 2.0. For the wav2vec 2.0 based methods, we train a sequence frame classifier based on the pretrained wav2vec 2.0 representations. We add an transformer layer and a linear layer for sequence frame classification at the top of the wav2vec 2.0 pretrained model. We keep the parameter of wav2vec 2.0 fixed and only train the sequence frame classifier.

5. Results and discussion

5.1. Results compared with the baselines

In Table 1, we compare the performance of different methods for SBD across four different languages on mTEDx dataset. We compute the f1 score of the frame-wise predictions for sentence boundary and get the average f1 score across the four tested languages among different methods.

We observe that the proposed GL-SSD achieves a better performance than other recent mainstream methods across all the four tested languages. The proposed GL-SSD outperforms the VAD-hybrid method and the wav2vec 2.0 based methods by 4.9 points and 2.5 points in average, respectively. We also observe that the wav2vec 2.0 based method perform better than the

[2] https://github.com/wiseman/py-webrtcvad

Table 1: *Results compared with the baselines*

SBD methods	Fr	Es	Pt	It	Avg
VAD-hybrid	0.733	0.691	0.681	0.662	0.692
Wav2vec 2.0	0.761	0.712	0.701	0.689	0.716
GL-SSD(ours)	**0.793**	**0.732**	**0.725**	**0.715**	**0.741**

Table 2: *Ablation results for different architectures*

configuration	Fr	Es	Avg
SR	0.761	0.712	0.737
SR + GSR	0.781	0.725	0.753
SR + LSR	0.773	0.718	0.746
SR + GSR + LSR (GL-SSD)	**0.793**	**0.732**	**0.763**

VAD-hybrid method, which demonstrate the importance of semantic informationfor SBD. The VAD-hybrid method achieve a lower f1 score because it only considers the acoustic silence but does not predict the semantic boundary. Due to the hybrid representation of the global-local speech styles and semantic information, the proposed method improves the performance for SBD by a large margin.

5.2. Ablation study

5.2.1. Respective effects of the global and local style representation

In order to investigate the respective additional improvement brought by the disentangled representation of global speech style and local speech style, we conduct an ablation study on the dataset of French and Spanish. We compared the results of four experimental setups as follows:

a). Semantic representation (SR). Keep only the semantic encoder and train the sequence classification model for SBD using only the semantic information, which is equal to the wav2vec 2.0 based method.

b). Semantic representation + Global style representation (SR+GSR). Incorporate the global speech style representation into the semantic representation, without the local style representation.

c). Semantic representation + Local style representation (SR+LSR). Incorporate the local speech style representation into the semantic representation, without the global style representation.

d). Semantic representation + Global style representation + Local style representation (SR+GSR+LSR). Incorporate both the representations of the global speech style and the local speech style into the semantic representation, which is the proposed GL-SSD.

Table 2 shows the results of the ablation study for the four experimental setups described above. We observe that both of the global style representation and the local style representation can bring additional improvement on the base of the semantic representation. The global style representation brings an improvement of 1.65 points and the local style representation brings an improvement of 0.9 points. When Combining the global style representation and the local style representation, the average improvement increases to 2.6 points. The results demonstrate that the hybrid representation of global style representation, the local style representation and the semantic representation performs better than each of them respectively. We also find that the global style representation brings more improvement than the local style representation. We consider the global speech style may probably play a more important role for sentence boundary detection than the local style to some degree.

In order to further investigate the relationship between the learned representation and the speech style, we project the learned embedding of the global codebook into 2D space by t-distributed stochastic neighbor embedding (TSNE) [28]. We randomly pick several speech talks of different speakers and extract their global style codes using the learned encoders and codebooks. We compute the mean and variance of the pause lengths according to the annotations. We find that for the speakers who have similar mean and variance of the pause lengths, the projections of their global style codes also stand close to each other.

Table 3: *Comparison results of vector-quanzatized representation and continuous representation*

configuration	Pt	It	Avg
continuous representation	0.712	0.701	0.707
vector-quantized representation	**0.725**	**0.715**	**0.720**

5.2.2. Effects of the vector-quantized representation

We further conduct a experiment to investigate the effect of the vector-quantized discrete representation compared to the continuous representation. Specifically, we remove the global codebook and the local codebook, only keep the global encoder and the local encoder. The output features of the global encoder and local encoder are fused directly with the semantic representation. Other experimental setups are the same as the proposed GL-SSD. The result is shown in Table 3.

From Table 3, we can obtain that the vector-quantized discrete representation can achieve a better performance compared to the continuous representation. We consider it is probably because of the generalization and robustness of the vector quantization for the self-reconstructed training. Similar conclusion is obtained in [29] which improves both robustness and generalization of vision models by the discrete adversarial training.

6. Conclusions

In this paper, we propose a robust method for sentence boundary detection in the speech stream by incorporating the disentangled representations of global speech style and local speech style into the semantic representation. Our method first adopts a self-supervised approach to disentangle the global and local speech styles by vector quantization from the raw speech. Then we use the hybrid representations of the learned disentangled global-local style representations and the semantic representation to train a sequence classification model for robust sentence boundary detection. The experimental results demonstrate the superior performance of the proposed method. In the ablation study, we further verified the respective improvements brought by the disentangled global style representation and the local style representation. We also verified the effectiveness of the vector-quantized discrete representation compared to the continuous representation. The proposed GL-SSD improves the performance of SBD by a large margin in different languages compared to the mainstream methods.

7. References

[1] J. Read, R. Dridan, S. Oepen, and L. J. Solberg, "Sentence boundary detection: A long solved problem?" in *International Conference on Computational Linguistics*, 2012.

[2] A. Y. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," *CoRR*, vol. abs/1412.5567, 2014. [Online]. Available: http://arxiv.org/abs/1412.5567

[3] P. Koehn, A. Axelrod, A. B. Mayne, C. Callison-Burch, and D. Talbot, "Edinburgh system description for the 2005 iwslt speech translation evaluation," *Proc International Workshop on Spoken Language Translation*, 2005.

[4] E. Matusov, P. Wilken, P. Bahar, J. Schamper, P. Golik, A. Zeyer, J. A. Silvestre-Cerdà, A. A. Martinez-Villaronga, H. Pesch, and J.-T. Peter, "Neural speech translation at apptek," in *International Workshop on Spoken Language Translation*, 2018.

[5] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly transcribe foreign speech," *CoRR*, vol. abs/1703.08581, 2017. [Online]. Available: http://arxiv.org/abs/1703.08581

[6] E. Matusov, G. Leusch, O. Bender, and H. Ney, "Evaluating machine translation output with automatic sentence segmentation," in *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA, Oct. 24-25 2005. [Online]. Available: https://aclanthology.org/2005.iwslt-1.19

[7] H. Moniz, F. Batista, H. Meinedo, A. Abad, I. Trancoso, A. I. Mata, and N. Mamede, "Prosodically-based automatic segmentation and punctuation," in *Proc. Speech Prosody 2010*, 2010, p. paper 910.

[8] T. Biron, D. Baum, D. Freche, N. Matalon, N. Ehrmann, E. Weinreb, D. Biron, and E. Moses, "Automatic detection of prosodic boundaries in spontaneous speech," *PLoS ONE*, vol. 16, 2021.

[9] E. Salesky, M. Wiesner, J. Bremerman, R. Cattoni, M. Negri, M. Turchi, D. W. Oard, and M. Post, "The multilingual tedx corpus for speech recognition and translation," in *Interspeech*, 2021.

[10] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 1, pp. 127–154, 2000, accessing Information in Spoken Audio. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639300000285

[11] M. Sinclair, P. Bell, A. Birch, and F. R. McInnes, "A semi-markov model for speech segmentation with an utterance-break prior," in *Interspeech*, 2014.

[12] W. Lu and H. T. Ng, "Better punctuation prediction with dynamic conditional random fields," in *Proceedings of the 2010 conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2010, pp. 177–186.

[13] T. Nguyen and S. Vogel, "Context-based arabic morphological analysis for machine translation," in *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, ser. CoNLL '08. USA: Association for Computational Linguistics, 2008, p. 135–142.

[14] C. Xu, L. Xie, G. Huang, X. Xiao, E. Chng, and H. Li, "A deep neural network approach for sentence boundary detection in broadcast news," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2887–2891, 01 2014.

[15] M. T. Diab, K. Hacioglu, and D. Jurafsky, "Automatic tagging of arabic text: From raw text to base phrase chunks," in *North American Chapter of the Association for Computational Linguistics*, 2004.

[16] E. Matusov, D. Hillard, M. Magimai-Doss, D. Hakkani-Tür, M. Ostendorf, and H. Ney, "Improving speech translation with automatic boundary prediction," in *Interspeech 2007*. ISCA, aug 2007. [Online]. Available: https://doi.org/10.21437%2Finterspeech.2007-644

[17] V. K. R. Sridhar, J. Chen, S. Bangalore, A. Ljolje, and R. Chengalvarayan, "Segmentation strategies for streaming speech translation," in *North American Chapter of the Association for Computational Linguistics*, 2013.

[18] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, pp. 1–3, 1999.

[19] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[20] H. Inaguma, B. Yan, S. Dalmia, P. Guo, J. Shi, K. Duh, and S. Watanabe, "ESPnet-ST IWSLT 2021 offline speech translation system," in *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*. Bangkok, Thailand (online): Association for Computational Linguistics, Aug. 2021, pp. 100–109. [Online]. Available: https://aclanthology.org/2021.iwslt-1.10

[21] M. Gaido, M. Negri, M. Cettolo, and M. Turchi, "Beyond voice activity detection: Hybrid audio segmentation for direct speech translation," in *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*. Trento, Italy: Association for Computational Linguistics, 12–13 Nov. 2021, pp. 55–62. [Online]. Available: https://aclanthology.org/2021.icnlsp-1.7

[22] M. Kunešová and M. Řezáčková, "Detection of prosodic boundaries in speech using wav2vec 2.0," in *Text, Speech, and Dialogue*, P. Sojka, A. Horák, I. Kopeček, and K. Pala, Eds. Cham: Springer International Publishing, 2022, pp. 377–388.

[23] G. I. Gállego, I. Tsiamas, C. Escolano, J. A. R. Fonollosa, and M. R. Costa-jussà, "End-to-end speech translation with pre-trained models and adapters: Upc at iwslt 2021," in *International Workshop on Spoken Language Translation*, 2021.

[24] I. Tsiamas, G. I. Gállego, J. A. R. Fonollosa, and M. R. Costa-jussà, "SHAS: Approaching optimal Segmentation for End-to-End Speech Translation," in *Proc. Interspeech 2022*, 2022, pp. 106–110.

[25] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6309–6318.

[26] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 30, p. 495–507, jan 2022. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3129994

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[28] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. nov, pp. 2579–2605, 2008, pagination: 27.

[29] X. Mao, Y. Chen, R. Duan, Y. Zhu, G. Qi, S. Ye, X. Li, R. Zhang, and H. Xue, "Enhance the visual representation via discrete adversarial training," *CoRR*, vol. abs/2209.07735, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2209.07735