Aligning Speech Segments Beyond Pure Semantics

Kevin Heffernan
Meta AI
kevinheffernan@meta.com

Artyom Kozhevnikov
Meta AI
artyomko@meta.com

Alexandre Mourachko
Meta AI
alexmourachko@meta.com

Loïc Barrault
Meta AI
loicbarrault@meta.com

Holger Schwenk
Meta AI
schwenk@meta.com

Abstract

Multilingual parallel data for speech-to-speech translation is scarce and expensive to create from scratch. This is all the more true for expressive speech translation, which aims at preserving not only the semantics, but also the overall prosody (e.g. style, emotion, rate-of-speech). Existing corpora contain speech utterances with the same meaning, yet the overall prosody is typically different, as human annotators are not tasked with reproducing these aspects, or crowed-sourced efforts do not specifically target this kind of alignment in priority. In this paper, we propose a novel alignment algorithm, which automatically forms pairs of speech segments aligned not only in meaning, but also in expressivity. In order to validate our approach, we train an expressive multilingual speech-to-speech translation system on the automatically aligned data. Our experiments show that in comparison to semantic-only approaches, expressively aligned data yields large improvements in source expressivity preservation (e.g. 43% uplift in speech rate preservation on average), while still maintaining content translation quality. In some scenarios, results also indicate that this alignment algorithm can outperform standard, semantic-focused approaches even on content translation quality.

1 Introduction

In traditional machine translation (MT), the underlying goal is to preserve the meaning of the source. However, more recently there have been efforts to develop expressive speech-to-speech translation systems (S2ST) (Seamless Communication et al., 2023), where the aim is to maintain not only the meaning, but also the expressivity of the source speech (e.g. tone, emotion, style, etc). While different parts of translation models can be pretrained in an unsupervised manner, large amounts of high-quality end-to-end data remains crucial to achieve the best performance. On one hand, human-curated parallel data for the text domain (bitexts), are freely available for several languages, for instance the well known Europarl (Koehn, 2005) or UN corpora (Ziemski et al., 2016). On the other hand, human created aligned speech-to-speech parallel data is a scarce resource. To complement existing speech parallel data, automatic alignment algorithms have evolved as an important technique to provide additional data for a large number of languages and domains (Duquenne et al., 2023a; Seamless Communication et al., 2023).

However, these speech-to-speech alignment algorithms search only for speech segments with the same semantics, totally disregarding expressive properties of the source and target speech. In this paper, we extend similarity-based speech-to-speech alignment with an expressive criterion. Our experiments show that the use of expressively aligned data substantially boosts the preservation of expressivity in a multilingual expressive S2ST system. In addition, we observe that in some cases the expressively aligned data also improves content translation quality on some test sets. The main contributions of this work are:

- We propose the first speech-to-speech alignment algorithm which aligns not only the semantics, but also the expressivity of the source and target speech;
- We applied this approach to a publicly available raw corpus, and aligned approximately 12 thousand hours of English speech in five languages (French, German, Italian, Chinese Mandarin, and Spanish). The resulting dataset and metadata can be found online[1];
- We validate our approach by training a multilingual and expressive S2ST on the aligned data, yielding uplifts in expressivity preser-

[1] https://github.com/facebookresearch/seamless_communication/blob/main/docs/expressive/seamless_align_expressive_README.md

Findings of the Association for Computational Linguistics: ACL 2024, pages 3626–3635

vation, while maintaining content translation quality.

2 Methodology

Following Duquenne et al. (2023a), we first preprocess the data by segmenting the speech signal into plausible segments using a Voice Activity Detector model, followed by language identification of each segment, and then subsequently encode each speech segment into a multilingual embedding space introduced by Duquenne et al. (2023c). Once all segments are encoded, we then perform a k-nearest-neighbor search in the embedding space using the FAISS library (Douze et al., 2024) which allows for efficient search at scale. We then calculate a margin-based score over each candidate neighbor which has been shown to yield alignments which have a similar meaning (Artetxe and Schwenk, 2019a). In this work, we use the ratio margin R, defined as:

$$R(x,y) = \frac{cos(x,y)}{\sum_{z \in NN_k(x)} \frac{cos(x,z)}{2k} + \sum_{z \in NN_k(y)} \frac{cos(y,z)}{2k}}$$

where x and y are the source and target sentences, $cos(x,y)$ is the cosine similarity of x and y in the multilingual embedding space, and $NN_k(x)$ denotes the k nearest neighbors of x.

Since the multilingual speech encoders are trained with a teacher-student method using a text-based encoder as the teacher, this forces the speech representation to focus on semantics, ignoring other elements of speech such as the prosody (Duquenne et al., 2023a). Therefore, the k-nearest-neighbors retrieved in the embedding space above are likely based on semantic characteristics only. However, it is possible that from the neighbors retrieved, that the prosodic characteristics of some may better preserve the source than others.

In order to capture such signals, in comparison to previous approaches such as Duquenne et al. (2023a) which choose a neighbor based on semantics only, we instead choose a neighbor which maximises a blend of both semantics: R, with another term related to the prosodic similarity of the speech segments: P. We define our expressive scoring function E as follows:

$$\mathrm{E}(x,y) = \alpha \cdot \mathrm{R}(x,y) + (1-\alpha) \cdot \mathrm{P}(x,y)$$

where α controls the trade-off between semantic accuracy (R) and prosody preservation (P).

Instead of modifying the nearest neighbor search, a more straight-forward strategy could involve filtering existing speech alignments such as SPEECHMATRIX (Duquenne et al., 2023a) with a prosodic scoring function. However, the volume of the resulting dataset would likely be drastically reduced as no explicit prosody-preservation goal was enforced to begin with during the nearest neighbor search (i.e. an expressively and semantically aligned neighbor would be chosen by chance).

3 Experiment

In order to validate the effectiveness of expressive alignments, we perform a controlled experiment where we align a set of raw monolingual speech data using both a semantic-only scoring function, and our expressive scoring function. We then subsequently train two expressive S2ST systems on the resulting alignments from both approaches, and evaluate the results.

Benchmark datasets. We evaluate our models on the FLEURS, MEXPRESSO, MDRAL benchmark datasets. FLEURS (Conneau et al., 2023) in a partially n-way speech dataset with alignments in 102 languages, which is an extension of the text-based FLoRes-101 dataset (Goyal et al., 2022). MEXPRESSO and MDRAL (Seamless Communication et al., 2023) contain English speech aligned in five different target languages: Spanish, French, Italian, German, and Chinese (Mandarin). MDRAL is an extension of the Dialogues Re-enacted Across Languages (DRAL) Corpus (Ward et al., 2023), which contains aligned fragments from spontaneous conversations. MEXPRESSO on the other hand contains scripted sentences which are then recorded in various expressive styles (e.g. happy, sad, confused etc).

Evaluation metrics. In order to ensure that the expressive translation systems maintain content translation quality, we measure ASR-BLEU by transcribing the generated target audio using a publicly available Whisper model[2], and then subsequently calculate a BLEU score using SacreBLEU[3].

For expressivity-based metrics we follow Seamless Communication et al. (2023). Firstly, we measure speaker style similarity by encoding the source

[2] https://huggingface.co/openai/whisper-large-v2

[3] 13a tokenizer

and target audios using a pre-trained WavLM-based speaker style encoder (Chen et al., 2022), and then calculate the speaker style similarity as the cosine between source and target (Le et al., 2023). As rhythmic patterns in speech are an important aspect of expressivity, we also calculate both the speech rate and pause alignment. The rate-of-speech is calculated by measuring the number of syllables spoken per second. We then report the Spearman correlation of the number of syllables spoken between the source and target speech[4]. In addition to the speech rate, the pause alignment captures how well silences are preserved between the source and translation. Silence was measured using the Silero VAD system (Silero Team, 2021). For both speech rate and pause alignment metrics, we used the Rhythmic Toolkit implementation (Seamless Communication et al., 2023).

Prosodic scoring function. We experiment with various prosodic scoring functions P based on different potential prosodic speech signals. w2v-BERT (Chung et al., 2021) is a large-scale cross-lingual speech representation, and various sub-layers of w2v-BERT have been shown to contain strong prosodic characteristics (Seamless Communication et al., 2023). We explore each sub-layer of w2v-BERT for prosodic signals. PRETSSEL (Seamless Communication et al., 2023) is an expressive unit-to-speech generator which contains an expressive speech encoder capable of generating expressivity embeddings from the source and target speech. SONAR Expressive (Duquenne et al., 2023b) is a zero-shot expressive speech-to-speech translation system. Similar to PRETSSEL, it contains an expressivity encoder which has an explicit knowledge of prosodic speech signals. Lastly, AUTOPCP (Seamless Communication et al., 2023) is a neural-based model which is trained to predict Prosodic Consistency Protocol (PCP) scores (Huang et al., 2023) for a pair of speech inputs. PCP scores are measured on a likert scale between 1 and 4 (with 4 being the highest possible score), and have been found to correlate with human judgements. For the embedding-based prosodic signals, $P = cos(x, y)$, while for AUTOPCP we use a unit normalized PCP score for each pair of source and target speech segment.

In order to determine the optimal prosodic scoring function, we measure the percentage of incorrect alignments when attempting to re-align the MEXPRESSO development set using our expressive scoring function E. As MEXPRESSO contains sentences repeated in different styles, this makes it a good choice of benchmark as we can not solely rely on a semantic-based algorithm. In other words, if each sentence in the benchmark contained a different meaning, we would not need any prosodic signal in order to attempt to re-align the dataset.

We performed a grid-search over all possibilities of both P and α[5], yielding AUTOPCP as the best overall prosodic signal. In-depth results for each prosodic signal are reported in Appendix E. As w2v-BERT can also contain important semantic information, we also experimented using this as the source of semantic signal (R), but it did not improve results.

Aligning audio. Following the selection of prosodic scoring function, we then began the alignment procedure. Starting from a large publicly available source of diverse raw audio data totalling approximately 3.9 million hours (Seamless Communication et al., 2023), we applied audio segmentation using the Silero VAD model (Silero Team, 2021), and then subsequently applied language identification on each segment (Seamless Communication et al., 2023). Each resulting segment was then encoded into the semantic-based multilingual embedding space (Duquenne et al., 2023c), before we performed k-nearest neighbor search[6] and then applied our expressive scoring function. For semantic-only alignments we set $\alpha = 1$ (i.e. no contribution from the prosodic scoring function). In total we aligned 11.9k hours of English source speech in five languages: Spanish, French, German, Italian, and Chinese (Mandarin). The resulting dataset and metadata can be found online[7], along with the code[8].

Model training. Following the alignment of source audios, we then trained two multilingual expressive S2ST models with the same architecture as SeamlessExpressive (Seamless Communication et al., 2023) on each alignment type separately (semantic and expressive). Additionally as the difference between a semantic and expressive

[4]For Chinese Mandarin, characters are treated as syllables

[5]We explored possible values for α in $\{\frac{k}{10} \mid k \in [0, 10]\}$

[6]We set k=16 for all experiments

[7]https://github.com/facebookresearch/seamless_communication/blob/main/docs/expressive/seamless_align_expressive_README.md

[8]https://github.com/facebookresearch/stopes/blob/main/website/docs/pipelines/expressive_alignments.md

Direction	Corpus	Alignment	ASR-BLEU ↑	Vocal Style Similarity ↑	Speech Rate ↑	Pause ↑
xxx→eng	Fleurs	Expressive	29.46	**0.37**	**0.39**	**0.46**
		Semantic	**29.60**	0.36	0.33	0.45
	mDRAL	Expressive	35.28	0.25	**0.35**	**0.25**
		Semantic	**36.57**	0.25	0.20	0.23
	mExpresso	Expressive	**31.27**	0.25	**0.39**	**0.34**
		Semantic	30.39	0.25	0.30	0.30
eng→xxx	Fleurs	Expressive	**18.11**	0.22	0.52	0.31
		Semantic	17.35	0.22	**0.53**	**0.32**
	mDRAL	Expressive	22.88	0.30	**0.33**	**0.20**
		Semantic	**24.22**	**0.31**	0.16	0.19
	mExpresso	Expressive	**20.96**	0.24	**0.44**	0.38
		Semantic	20.33	0.24	0.34	0.38

Table 1: Model evaluation results (averaged over all languages) for both xxx→eng and eng→xxx directions.

alignment is a choice of neighbor in a shared knn space, the number of resulting alignments from both approaches is the same, which also controls for performance differences due to dataset size. We ensured that none of the aligned data was previously seen in any pre-training. The architecture and hyperparameters of both models was identical, and each was trained for the same number of steps. More in-depth details can be found in Appendix C.

4 Results

Model evaluation results are shown in Table 1. On average, both the speech rate and pause alignment expressive metrics improve in relation to the model trained on semantic alignments only. On average we see a 43% relative improvement in speech rate, while still maintaining content translation quality (-0.08 BLEU on average). In particular, on the MDRAL benchmark we see relative speech rate improvements of 106% (0.16→0.33) on eng→xxx, and 75% (0.20→0.35) on xxx→eng. For content translation quality, in three instances we see that on average the model trained on expressive alignments even outperforms the semantic model. For example, we see a +0.88 BLEU improvement on MEXPRESSO (xxx→eng). We observed that there was very little difference between both models on vocal style similarity. However, this is perhaps partially due to the fact that the PRETSSEL unit-to-speech component of the S2ST model which mostly affects this metric was pre-trained beforehand (Seamless Communication et al., 2023), suggesting this component had previously converged. A more detailed breakdown showing expanded results per language for each dataset, along with supplemental ASR-COMET scores can be seen in Appendix D.

5 Related Work

Research on aligning texts was initially based on document meta-information (Resnik, 1999), cross-lingual document retrieval (Munteanu and Marcu, 2005) or machine translation and information retrieval (Abdul-Rauf and Schwenk, 2009; Bouamor and Sajjad, 2018). However, many current alignment techniques are based on a similarity measure in a multilingual embedding space (Artetxe and Schwenk, 2019b; Feng et al., 2020). One such technique for aligning text was introduced by Schwenk et al. (2019), which uses a margin-based measure of similarity (Artetxe and Schwenk, 2019a) in order to determine the candidacy of potential alignments. This technique was then extended to the speech modality (Duquenne et al., 2021), where it was used to create high-quality speech-to-speech aligned data such as SPEECHMATRIX (Duquenne et al., 2023a) which covers seventeen languages, and SEAMLESSALIGN (Seamless Communication et al., 2023) which provides a total of 585k hours in 95 languages. Both corpora consider only semantics during the alignment procedure.

6 Conclusion

We introduce the first speech-to-speech alignment method which can align speech not only in terms of semantics, but also the expressivity. We validate our method by performing large-scale speech-to-speech alignment, and train an expressive S2ST model on the resulting data. Our results show that expressive alignments can further boost the capability of expressive models, where such speech-to-speech data is extremely scarce, and in some instances even improve content translation quality.

7 limitations

We highlight three limitations of our work. The first is that we only expressively align English with five other high-resource languages: Spanish, French, Italian, German, and Chinese (Mandarin). Given the scarcity of such expressive speech data, it would be hugely beneficial for the community to cover more mid- and perhaps even some low-resource languages, as these are at risk of underexposure. However, as our expressive benchmark dataset used to tune our alignment algorithm (mExpresso) currently only supports these high-resource languages, this would require additional annotation efforts. Secondly, we experiment with two popular multilingual embedding spaces from Duquenne et al. (2023c) and Chung et al. (2021). However, there are other representations which would be interesting to explore which may help retrieve even better k-nearest-neighbors, and perhaps include some with even more prosidic preservation of the source speech, resulting in higher quality expressive alignments. Finally, we rely on automatic-based metrics in order to evaluate our models. However these do not give us a perfect assessment, and a human-based evaluation would yield more accurate results.

8 Acknowledgments

We want to extend our gratitude to those who made this work possible. To Peng-Jen Chen, Ilia Kulikov, Yilin Yang, Hongyu Gong, Yu-An Chung, Ann Lee, and Juan Pino who carried out modeling research on expressive translation and integrated our algorithm in their workflow. To Igor Tufanov, who helped with our first expressivity score prototype. And finally to David Dale and Benjamin Peloquin who developed AutoPCP and simplified its use in our expressive alignment algorithm.

References

Sadaf Abdul-Rauf and Holger Schwenk. 2009. On the Use of Comparable Corpora to Improve SMT performance. In *EACL*, pages 16–23.

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings. In *ACL*.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *TACL*, pages 597–610.

Houda Bouamor and Hassan Sajjad. 2018. H2@BUCC18: Parallel Sentence Extraction from Comparable Corpora Using Multilingual Sentence Embeddings. In *BUCC*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250. IEEE.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

Paul-Ambroise Duquenne, Hongyu Gong, Ning Dong, Jingfei Du, Ann Lee, Vedanuj Goswami, Changhan Wang, Juan Pino, Benoit Sagot, and Holger Schwenk. 2023a. SpeechMatrix: A large-scale mined corpus of multilingual speech-to-speech translations. In *ACL*, pages 16251–16269.

Paul-Ambroise Duquenne, Hongyu Gong, and Holger Schwenk. 2021. Multimodal and multilingual embeddings for large-scale speech mining. *Advances in Neural Information Processing Systems*, 34.

Paul-Ambroise Duquenne, Kevin Heffernan, Alexandre Mourachko, Benoît Sagot, and Holger Schwenk. 2023b. Sonar expressive: Zero-shot expressive speech-to-speech translation.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoit Sagot. 2023c. SONAR: sentence-level multimodal and language-agnostic representations.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Wen-Chin Huang, Benjamin Peloquin, Justine Kao, Changhan Wang, Hongyu Gong, Elizabeth Salesky, Yossi Adi, Ann Lee, and Peng-Jen Chen. 2023. A holistic cascade system, benchmark, and human evaluation protocol for expressive speech-to-speech translation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. *arXiv preprint arXiv:2306.15687*.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4):477–504.

Philip Resnik. 1999. Mining the Web for Bilingual Text. In *ACL*.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual expressive and streaming speech translation.

Silero Team. 2021. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. `https://github.com/snakers4/silero-vad`.

Nigel G. Ward, Jonathan E. Avila, Emilia Rivas, and Divette Marco. 2023. Dialogs re-enacted across languages.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1.0. In *LREC*.

A Number of source hours per benchmark dataset

	FLEURS		MEXPRESSO		MDRAL	
	dev	test	dev	test	dev	test
cmn	1.27	3.07	3.51	6.40	0.35	0.22
deu	1.26	3.15	4.85	7.21	0.83	0.92
fra	0.80	1.95	5.31	6.82	0.50	0.24
ita	1.55	3.52	5.86	6.64	0.68	0.99
spa	1.35	3.09	5.20	6.94	0.46	0.32

B Hours of automatically aligned audio per language

	Aligned hours
French	4,376
German	2,122
Italian	1,118
Mandarin	116
Spanish	4,242
Total	11,974

C Hyperparameters for S2ST system

text-to-unit encoder layers	4
text-to-unit encoder embed dim	1024
decoder layers	24
decoder attention heads	16
decoder embed dim	1024
decoder FFNN embed dim	8192
optimiser	Adam
adam betas	(0.9, 0.98)
learning rate	3e-5
dropout	0.1

D Detailed model evaluation results

D.1 xxx→eng direction

Corpus	Alignment	xxx→eng	ASR-BLEU	ASR-COMET	Vocal Style Similarity	Speech Rate	Pause
Fleurs	Expressive	cmn	21.21	0.80	0.36	0.10	0.45
		deu	37.64	0.82	0.38	0.65	0.45
		fra	33.60	0.81	0.37	0.36	0.45
		ita	28.11	0.82	0.37	0.43	0.47
		spa	26.75	0.81	0.35	0.41	0.48
	Semantic	cmn	21.86	0.80	0.35	0.09	0.44
		deu	37.82	0.82	0.38	0.53	0.44
		fra	33.54	0.81	0.37	0.33	0.47
		ita	28.24	0.82	0.37	0.31	0.44
		spa	26.54	0.81	0.35	0.37	0.47
mDRAL	Expressive	cmn	22.69	0.76	0.24	0.19	0.06
		deu	38.45	0.83	0.34	0.45	0.42
		fra	34.91	0.84	0.21	0.34	0.30
		ita	31.82	0.79	0.23	0.35	0.32
		spa	48.53	0.84	0.25	0.40	0.17
	Semantic	cmn	24.00	0.77	0.25	0.06	0.04
		deu	39.88	0.84	0.34	0.28	0.39
		fra	38.92	0.85	0.22	0.26	0.26
		ita	32.61	0.80	0.22	0.18	0.29
		spa	47.43	0.85	0.25	0.22	0.17
mExpresso	Expressive	cmn	20.79	0.75	0.27	0.42	0.29
		deu	29.01	0.75	0.28	0.46	0.33
		fra	30.78	0.77	0.21	0.36	0.39
		ita	33.19	0.76	0.26	0.34	0.35
		spa	42.57	0.80	0.25	0.36	0.36
	Semantic	cmn	20.81	0.74	0.26	0.33	0.21
		deu	27.61	0.73	0.27	0.38	0.29
		fra	30.09	0.75	0.20	0.30	0.33
		ita	31.82	0.75	0.25	0.23	0.32
		spa	41.64	0.79	0.24	0.26	0.32

D.2 eng→xxx direction

Corpus	Alignment	eng→xxx	ASR-BLEU	ASR-COMET	Vocal Style Similarity	Speech Rate	Pause
Fleurs	Expressive	cmn	0.26	0.30	0.20	0.48	0.24
		deu	19.47	0.61	0.21	0.52	0.32
		fra	32.63	0.66	0.22	0.50	0.33
		ita	18.51	0.66	0.22	0.55	0.34
		spa	19.69	0.71	0.24	0.56	0.34
	Semantic	cmn	0.26	0.31	0.19	0.47	0.24
		deu	17.39	0.57	0.21	0.53	0.35
		fra	31.76	0.66	0.21	0.54	0.34
		ita	17.87	0.66	0.22	0.57	0.35
		spa	19.45	0.71	0.24	0.55	0.35
mDRAL	Expressive	cmn	2.41	0.49	0.26	0.13	0.12
		deu	18.54	0.65	0.41	0.55	0.29
		fra	29.08	0.77	0.27	0.24	0.13
		ita	26.79	0.80	0.30	0.32	0.30
		spa	37.59	0.84	0.28	0.38	0.17
	Semantic	cmn	2.70	0.48	0.26	0.10	0.15
		deu	21.29	0.68	0.42	0.25	0.35
		fra	31.06	0.79	0.27	0.01	0.06
		ita	27.15	0.79	0.30	0.16	0.29
		spa	38.91	0.84	0.29	0.25	0.12
mExpresso	Expressive	cmn	2.03	0.43	0.21	0.29	0.30
		deu	15.85	0.63	0.27	0.52	0.38
		fra	26.02	0.67	0.24	0.45	0.42
		ita	26.01	0.74	0.25	0.46	0.40
		spa	34.87	0.77	0.24	0.48	0.42
	Semantic	cmn	2.03	0.43	0.21	0.26	0.32
		deu	13.46	0.60	0.27	0.41	0.38
		fra	25.02	0.65	0.24	0.34	0.41
		ita	26.09	0.73	0.25	0.33	0.38
		spa	35.06	0.77	0.24	0.36	0.40

E Detailed alignment error rate on MEXPRESSO using expressive scoring function (E)

Direction	Prosodic signal	spa	fra	ita	deu	cmn	average
xxx→eng	PRETSSEL	61.02	71.21	72.95	68.88	70.94	69.00
	SONAR Expressive	58.79	72.69	66.70	67.68	64.98	66.17
	w2v-bert-L0	63.80	74.42	69.79	73.50	65.76	69.45
	w2v-bert-L1	58.70	72.12	64.98	70.45	63.02	65.85
	w2v-bert-L2	56.77	70.81	64.38	69.21	62.66	64.77
	w2v-bert-L3	56.84	70.01	63.69	66.37	62.78	63.94
	w2v-bert-L4	56.00	69.49	61.74	64.31	61.72	62.65
	w2v-bert-L5	55.42	69.40	61.73	62.10	60.92	61.91
	w2v-bert-L6	53.40	68.58	60.82	59.93	60.42	60.63
	w2v-bert-L7	49.92	66.13	57.78	57.32	58.60	57.95
	w2v-bert-L8	49.82	65.87	57.68	58.17	59.22	58.15
	w2v-bert-L9	49.48	65.26	57.56	58.02	58.31	57.73
	w2v-bert-L10	48.95	65.01	57.21	57.70	58.31	57.44
	w2v-bert-L11	49.11	65.43	57.19	57.95	59.48	57.83
	w2v-bert-L12	49.62	65.76	57.02	58.00	59.80	58.04
	w2v-bert-L13	50.10	65.90	57.63	57.77	61.20	58.52
	w2v-bert-L14	49.08	65.29	57.37	56.99	60.72	57.89
	w2v-bert-L15	48.73	65.71	57.07	55.31	61.16	57.60
	w2v-bert-L16	48.16	65.95	56.05	53.58	60.58	56.86
	w2v-bert-L17	48.52	65.48	56.85	53.74	59.26	56.77
	w2v-bert-L18	49.60	66.13	57.87	55.33	59.50	57.69
	w2v-bert-L19	51.50	66.37	59.31	55.45	59.26	58.38
	w2v-bert-L20	54.40	67.42	62.30	60.25	63.66	61.61
	w2v-bert-L21	61.06	73.89	66.30	64.36	65.14	66.15
	w2v-bert-L22	84.86	85.72	84.35	83.67	82.59	84.24
	w2v-bert-L23	84.86	85.70	84.36	83.69	82.55	84.23
	AUTOPCP	**47.64**	**62.9**	**53.94**	**55.9**	**58.09**	**55.69**
	Semantic baseline ($\alpha = 1$)	84.86	85.53	84.38	83.81	82.67	84.25

Direction	Prosodic signal	spa	fra	ita	deu	cmn	average
eng→xxx	PRETSSEL	65.31	75.65	71.63	72.41	71.72	71.34
	SONAR Expressive	61.34	74.12	66.31	69.53	65.32	67.32
	w2v-bert-L0	63.11	74.97	70.15	73.26	69.18	70.13
	w2v-bert-L1	61.20	73.60	68.94	71.83	69.10	68.93
	w2v-bert-L2	61.53	73.34	68.57	71.43	69.02	68.78
	w2v-bert-L3	61.18	73.11	68.00	70.85	69.00	68.43
	w2v-bert-L4	60.71	72.95	67.65	69.51	68.94	67.95
	w2v-bert-L5	58.41	71.40	65.74	68.74	67.32	66.32
	w2v-bert-L6	56.03	70.50	64.77	68.10	67.32	65.34
	w2v-bert-L7	52.61	68.44	62.00	66.81	67.04	63.38
	w2v-bert-L8	53.35	69.58	63.08	67.92	67.98	64.38
	w2v-bert-L9	53.68	70.17	64.16	68.88	67.42	64.86
	w2v-bert-L10	53.87	70.59	64.33	69.30	67.74	65.17
	w2v-bert-L11	54.33	71.25	65.22	69.75	67.68	65.65
	w2v-bert-L12	54.49	72.17	65.86	70.16	67.48	66.03
	w2v-bert-L13	55.26	72.07	65.67	69.63	67.20	65.97
	w2v-bert-L14	53.50	71.54	63.48	68.03	65.88	64.49
	w2v-bert-L15	51.54	70.38	61.36	66.34	65.16	62.96
	w2v-bert-L16	49.80	68.91	59.09	63.60	63.06	60.89
	w2v-bert-L17	48.95	68.81	58.81	62.60	62.16	60.27
	w2v-bert-L18	51.54	69.99	60.22	63.82	64.08	61.93
	w2v-bert-L19	54.96	71.53	62.18	67.29	66.70	64.53
	w2v-bert-L20	60.04	73.49	64.94	70.66	69.20	67.67
	w2v-bert-L21	65.20	76.68	69.61	74.83	71.24	71.51
	w2v-bert-L22	84.52	83.93	83.69	84.13	82.39	83.73
	w2v-bert-L23	84.52	83.94	83.69	84.14	82.39	83.74
	AUTOPCP	**49.60**	**63.88**	**54.40**	**56.41**	**57.71**	**56.40**
	Semantic baseline ($\alpha = 1$)	84.56	83.98	83.65	84.16	82.41	83.75