Segmentation Strategies

Mihai-Andrei Tudor
Faculty of Data Science and Artificial Intelligence, Maastricht University
m.tudor@student.maastrichtuniversity.nl

Abstract

The speech translation domain saw two main research paradigms, the cascaded approach, and the recently emerged end-to-end technique. However, given the undeniable importance of speech segmentation and the emergence of several automated segmentation software, the outcomes of employing different segmentation strategies on the end-to-end model compared to the cascaded model are still under-studied.

The study presented in this work has two main focuses: firstly, investigating three modern automatic speech segmentation toolkits. Secondly, and more importantly, the paper experiments by employing an end-to-end and a cascaded speech translation model. The paper will analyze the consequences of providing automatic segmentation to an end-to-end and a cascaded model. The conducted work proves that recently released automatic speech detection software such as Voxseg can deliver robust results in the end-to-end system and therefore surpass the other software. However, the paper also indicates that applying the same segmentations to the cascaded model often leads to a different outcome, making other techniques more suitable in various circumstances.

1 Introduction

There has been a dramatic improvement in speech translation in the last few years. While this progress is also attributed to immense corpora recently being released to the public, the leading cause for this swift advance is the adoption of different deep learning techniques.

Being released in 2016, a promising new method to translate speech directly became more and more relevant in academic circles - end-to-end approaches (Berard, Besacier, Kocabiyikoglu, & Pietquin, 2018). Despite representing a significant milestone for the speech translation (ST) community, the first end-to-end model had apparent impediments in the short datasets used for training and the extensive use of synthetic speech augmented data. However, as larger quantities of corpora were released, end-to-end methods started exceeding cascaded approaches in specific yet realistic circumstances (Sperber & Paulik, 2020).

One of these corpora includes thousands of recordings of TEDx Talks in eight source languages (Salesky et al., 2021). While referring to these specific corpora, evident aspects such as the varying microphone characteristics, acoustic conditions in the talk room might degrade inputs for acoustic modeling, making it a complicated assignment despite the extents of the recently released corpora. Furthermore, each TEDx talk contains large parts with no speech activity apart from the aspects mentioned above. Therefore, skipping non-speech segments, on the one hand, decreases the latency in the automatic speech recognition process. On the other hand, segmentation also improves accuracy since insertions due to incorrectly labeled background voices from the public or music is harmful to the translation quality, which is particularly useful considering that hundreds of people could potentially be in the room.

While machine translation (MT) can be performed by simply conducting a word-by-word translation, that procedure alone cannot deliver reasonable translation hypotheses because recognizing whole sentences and their equivalents in the target language are required. This is why voice activity detection (VAD) is vital in identifying the boundary markers for each sentence as accurately as possible. In this regard, deep learning also led to modern VAD software being established in the significant advancement of speech translation, leading to the earlier mentioned declines in latency and better translation accuracy.

This paper will use three modern VADs and involve different segmentation strategies for each one. All resulting segmentations will be inputs for already existing cascaded and end-to-end translation models. Finally, the translation hypothesis will be evaluated. While the paper also compares the three VADs and the translation accuracies obtained with either of the segmentations, the primary goal of this paper is to investigate the outcome of employing the segmentation strategies on a cascaded multilingual translation system compared to an end-to-end one.

2 Related Work

Firstly, regarding the automatic speech detection software, articles submitting new hybrid architecture VADs such as Voxseg (Wilkinson & Niesler, 2021) train and evaluate its proposed software on in-the-wild corpora composed of Youtube videos from the AVA-Speech dataset (Chaudhuri et al., 2018). One distinctive characteristic about this corpus is that it only contains 45 hours of labeled data, half of which is non-speech.

Other earlier released VADs only apply more conventional yet larger convolutional neural networks (Doukhan, Lechapt, Evrard, & Carrive, 2018). However, older such VADs software like inaSpeechSegemnter is yet widespread, and the paper introducing the toolkit evaluates it on French videos corpora (Giraudel et al., 2012). Moreover, unlike the previous VAD, inaSpeechSegmenter has been trained on a collection of three distinct datasets, totalizing around 70 hours of recordings, mainly consisting of speech labels. These three corpora will be briefly described in Section 4.2. Presumably, due to the benefits and weaknesses of each VAD as reported by its developers and also because of the emergence of new ones every year, the comparison of such software has not been assessed in the recent specialized literature.

Secondly, all experiments were conducted by employing two already existing state-of-the-art models representing two paradigms in the domain of speech translation. Both systems, the end-to-end and the on based on the cascaded approach, were submitted by Maastricht University for IWSLT 2021 (Liu & Niehues, 2021). The work displayed in this paper did not implicate the models' conceptualization, coding, or training procedures. Likewise, the author of this paper was not involved in the IWSLT campaign submission. Thus, all tests and comparisons shown in this paper were possible due to the earlier work submitted by Maastricht University for IWSLT 2021 and its participants.

Lastly, concerning the segmentation strategies, a previous study analyzes segmentation to maximize the number of segments while still maximizing translation accuracy (Rangarajan Sridhar, Chen, Bangalore, Ljolje, & Chengalvarayan, 2013). For this, the paper performed experiments with several linguistic and non-linguistic strategies for text segmentation before translation (e.g., conjunction-separated sentence chunks vs. comma-separated sentence chunks). While the segmentation strategies employed in this paper are entirely different, the same evaluation method was employed.

This paper will employ automatic segmentation software, including the two VADs mentioned above, on a different audio corpus than those previously used for the toolkits' training and evaluation procedures. To obtain hypotheses, the resulting segmentations will be inputs for the aforementioned pre-trained speech translation models. Ultimately, the translation accuracy of each of the hypotheses will be evaluated at the talk level through the BLEU metric, similarly to one of the evaluation methods used in one of the cited related work, overlooking the latency evaluation nonetheless.

3 Automatic Segmentation for Speech Translation

3.1 IWSLT 2021 Multilingual Speech Translation System

The two employed multilingual speech translation systems were previously trained in English, Spanish, French, Italian, and Portuguese. Hence plenty of other tests not presented in this paper could have been additionally performed (e.g., French-English). Furthermore, while a comprehensive description of the models can be found in the cited paper, in summary, it states that pseudo-labels were first incorporated in the end-to-end system's training process (E3). Then the model was enhanced even further by using the same pseudo-labeling method on the supervised directions, resulting in the vastly improved E4 model used in this paper. Finally, the same end-to-end (E4) model was again used to feed the ASR transcriptions in the context of the cascaded speech translation system (E4+M1).

3.2 Corpus Statistics

As earlier mentioned, the most extensive multilingual speech translation corpus is mTEDX (Salesky et al., 2021). Each TED recordings has been manually segmented, transcripted, and translated by volunteers. In addition, the paper noted an average length of 10 minutes among all the audio recordings composing the corpora. Despite this, due to the splitting method

(Section 2.4. of the cited paper), we can see in Table 1 that specific source languages such as Portuguese have moderately lengthier test sets (i.e., 121 minutes test set vs. 109 minutes on validation set).

This work focused mainly on tests performed on the Italian-English, Portuguese-Spanish, Portuguese-English, Spanish-English, and Spanish-Portuguese language pairs from the test and validation corpora. The paper that introduced the utilized end-to-end and cascaded systems showed better results by using the end-to-end approach for most pairs, including Italian-English and Portuguese-Spanish when employing manual segmentation of the recordings (Liu & Niehues, 2021). On the other hand, there were also source-target pairs where the cascaded model reached better scores (e.g., Spanish-English and Portuguese-English). This observation, combined with the length of test-validation sets, led to choosing these specific language pairs for tests to provide robustness in the results.

Source	Target (minutes, #utts.)					
	EN			ES		
	train	test	valid	train	test	valid
PT	3681, 30855	121, 1022	109, 1013	1430, 11499	121, 1020	109, 1013
IT	3322, 24576	132, 999	131,931	330, 2261	132, 979	131, 931
ES	4661, 36263	123, 996	117,905			

Table 1: Data amount of speech translation in the training, test, and validation sets of mTEDx.

3.3 Preprocessing

For the audio data, all the original FLAC (Free Lossless Audio Codec) files in the test and validation sets first had to be converted into waveform files, downsampled from 48000Hz to 16000Hz and then converted from stereo to mono by mixing the two channels. All these procedures were performed by using the Sox library (R. M. Bittner & Bello, 2016).

4 Experiments

Compared to the early systems back in the 1960s that only involved a straightforward threshold to the signal's energy to detect the presence of speech (Bullington & Fraser, 1959), modern VAD toolkits use various methods of deep-learning and Hidden Markov models. However, despite innovative techniques being used, each software still has similar adjustable parameters that hugely impact each segment's boundary markers and, therefore, the resulting translation.

4.1 Voxseg VAD (Wilkinson & Niesler, 2021)

The first studied toolkit is presented in (Wilkinson & Niesler, 2021). The paper released in 2021 presents a unique hybrid architecture for VAD tasks. The architecture incorporates both convolutional neural network (CNN) and bidirectional long short-term memory (BiLSTM) layers trained in an end-to-end method. Apart from the structure, one other distinctive characteristic was the promise that the model focused on optimizing the architecture to provide robust performance in challenging, noisy conditions.

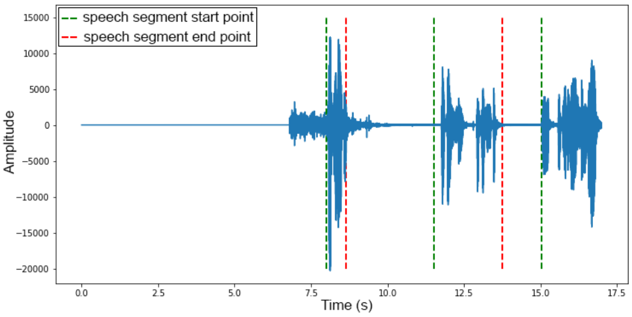

Figure 1: Illustration representing the first seventeen seconds of segmentation executed by the Voxseg VAD (Wilkinson & Niesler, 2021) when using the manual speech threshold value 0.5 (i.e., -s 0.5) on one of the Italian test files - 3rKh3t7NCzw.

The following table indicates the previously mentioned evenly split labels between speech and non-speech segments, unlike other speech-heavy datasets. More notably, the average duration of approximately three seconds in speech segments should be observed. This aspect is essential in underlining that the VAD will be tested in different conditions than the Youtube recordings used in training. For context, the average durations in the manual segmentations in the mTEDX datasets range from approximately six to seven seconds.

Label	Time	Average duration
Clean speech	14.55%	2.97 sec
Speech + music	13.46%	3.43 sec
Speech + noise	24.32%	3.28 sec
No speech	47.68%	3.68 sec

Table 2: Aggregate statistics over the Ava-Speech dataset (Chaudhuri et al., 2018).

This toolkit's most commonly used parameters are the speech vs. non-speech decision threshold (i.e., -s) accepting float values between 0 and 1 and the median filtering to smooth the output (i.e., -f). While various decision thresholds were tested in the experiments section, the second parameter remained the same, utilizing the manual kernel size of 1.

4.2 inaSpeechSegmenter VAD (Doukhan et al., 2018)

The second software used for experiments was preferred for two main reasons: it is the winner of MIREX 2018, and, unlike the other two toolkits, it incorporates an additional module of the framework that labels the resultant segments according to the corresponding gender (Doukhan et al., 2018). Although particularly interesting and hence making it one of the most popular VAD packages on Github, this element has not been evaluated in this paper. Instead, the experiments ignore the gender and use the output to create segments that can be utilized for computing MFCC features using Kaldi (Povey et al., 2011).

As previously mentioned, the training collection used for inaSpeechSegmenter is impressive compared to the 45 hours of recordings used for Voxseg. However, most data comes from audio tracks, a somewhat different context than TEDx Talks.

- GTZAN: 1000 audio tracks each 30 seconds long from 10 music genres (Sturm, 2014)
- Schreirer-Slang: collection of some 240 15-second extracts collected from the radio (Scheirer & Slaney, 1997)
- Musan: a total of 60 hours of speech, 42 hours of music and a total of six hours of noise (Snyder, Chen, & Povey, 2015)

Like the speech threshold present in the Voxseg VAD, a simple energy threshold called energy ratio is used to discard frames associated with low energy. This parameter is the only flag adjusted throughout the experiments to discover a proper segmentation given the TEDx corpus. The manual threshold is 0.03, and, given that it takes any float values from 0 to 1, the optimal value cannot be calculated without apt computing power. Nevertheless, six different parameters will be shown, ranging from 0.01 to 0.20. The Results section will show that the BLEU scores given by the end-to-end multilingual translation model will eventually reach a cutoff point by increasing the threshold (and boosting the number of segments). Thus the parameter shall not be incremented further given the reducing translation scores.

4.3 py-webrtcvad VAD (Wiseman, 2020)

Ultimately, the paper conducted experiments on py-webrtcvad (Wiseman, 2020), a python wrapper for the VAD that Google developed for the WEBRTC project (Google, 2011) that numerous other companies such as Mozilla and Opera support. The GitHub repository suggests that the VAD has one major issue, "that speech signal is considered alone, and the methods for arbitrary audio signal recognition are in a pretty initial stage." Also, overlapping audio signals is in the initial stage. While this aspect seems like a significant matter compared to the Voxseg toolkit, in an organized circumstance such as a TEDx presentation, at least in theory, it did not emerge as a substantial problem.

labels	start	stop
music	0.0	22.48
noEnergy	22.48	29.080002
noise	29.080002	52.80004
male	52.80004	68.20006

Table 3: Exemplification of the comma-separated values (CSV) output format of inaSpeechSegmenter toolkit.

A brief look into the algorithm reveals a padded, sliding window over the audio files. When more than 90% of the frames in that specific window are labeled as voiced, it triggers the start of speech segments, terminating it when less than 90% of the frames in the window are labeled so by the VAD. This type of algorithm leads to another potential adjustable parameter that has not been open in the earlier toolkits - the padding duration. For the experiments conducted in this paper, we will be using the same padding used in the provided test script (i.e., 30 milliseconds). The computational reasons and the dissimilarity of this parameter compared to the available ones in the previous toolkits justify using the same 30 milliseconds padding for all tests.

Apart from the padding window, one other modifiable flag appears - aggressiveness. The parameter modifies how aggressively to filter out non-speech. 0 being the least aggressive and three the most aggressive. However, unlike the previous software accepting float values, this parameter only takes natural numbers as inputs, allowing a maximum number of four segmentation strategies per corpus.

5 Results

Before looking at the BLEU scores reached with the three VADs, it is still essential to show the complete talk-level BLEU scores achieved using the manually annotated segmentation. The test and validation set evaluations on the talk level are presented in Table 4.

Out of the ten tested language pairs, manual segmentations led to the end-to-end multilingual translation model obtaining better BLEU scores on six occasions. Furthermore, in all four tests that involved Italian as a source language, the end-to-end system outperformed the cascaded one by a large margin (e.g., differences of

ST direction	E2E (E4)		Cascaded (E4+M1)		E2E vs. Cascaded percentage difference
	Segments count	BLEU	Segments count	BLEU	
PT-ES_test_manual_seg	1022	39.3	1022	39.2	0.3%
PT-ES_valid_manual_seg	1013	44.5	1013	43.6	2.0%
IT-EN_test_manual_seg	979	31.0	979	29.3	5.6%
IT-EN_valid_manual_seg	931	27.9	931	26.8	4.0%
IT-ES_test_manual_seg	979	35.4	979	33.2	6.4%
IT-ES_valid_manual_seg	931	36.1	931	33.3	8.1%
ES-EN_test_manual_seg	996	41.3	996	43.0	4.0%
ES-EN_valid_manual_seg	965	37.6	965	38.9	3.4%
PT-EN_test_manual_seg	1022	34.2	1022	34.6	1.2%
PT-EN_valid_manual_seg	1013	38.1	1013	39.3	3.1%

Table 4: BLEU scores at the talk level and segments count for the Portuguese-Spanish, Italian-English, Italian-Spanish, Spanish-English, and Portuguese-English language pairs when using the original segmentations contained in the mTEDx test and validation datasets (Salesky et al., 2021).

5.6% and 6.4% on the test sets). In contrast, the Portuguese corpus led to two contrasting results depending on the target language. On the one hand, the end-to-end model achieved negligibly better results when Spanish was the target language (i.e., 0.3% and 2.0% on the test and validation set, respectively) and, on the other hand, inferior results compared to the cascaded model when translating to English (i.e., 1.2% and 3.1%).

Concerning the number of manually annotated segments, the three corpora (Portuguese, Italian and Spanish) had an average of approximately 1000 segments on the test set and 970 on the validation set.

5.1 Automatic Segmentations Employing the End-To-End Model

When segmentation was performed using the VADs, the Voxseg software surpassed the others on six occasions (Table 5 and Table 7). Four of these tests involve the usage of Portuguese as a source language (i.e., Portuguese-Spanish and Portuguese-English on test and validation sets), and two of the four conducted tests implicating the Italian TEDx Talks. A speech vs. non-speech threshold of -s 0.95 led to four of these best-performing test cases. Still subject to these six specific tests, there were two cases in which inaSpeechSegmenter could provide an equally accurate translation when employing the end-to-end approach. However, it would require more segments than Voxseg to reach the same translation accuracy in both circumstances.

Despite employing fewer tests due to a limited adjustment in its parameter, the Webrtcvad VAD topped in all the remaining four tests, predominantly in tests implicating Spanish as a source language. While other best-found segmentations showed significant differences in terms of translation quality compared to the manual segmentations (e.g., 17.7% on the Portuguese-Spanish test set), the Webrtcvad VAD was able to approach the closest when translating the earlier mentioned Spanish recordings into English (i.e., 6.5%). Moreover, Webrtcvad exceeded the other VADs on these four occasions by producing significantly fewer segments than the tests in which Voxseg achieved better results (e.g., only 11.4% more segments on the Spanish-English test set as compared to 23.5% more segments in the Portuguese-Spanish for Voxseg).

Language pair	Best segmentation	Segments count	Segments Difference*	BLEU score Difference*
pt-es_test	voxseg -s 0.90	1294	23.5%	17.7%
pt-es_valid	voxseg -s 0.95	1139	11.7%	18.1%
it-en_test	voxseg -s 0.95	1223	22.2%	14.9%
it-en_valid	webrtcvad -p 2	1075	14.4%	9.8%
it-es_test	voxseg -s 0.95 inaspeech -r 0.15 inaspeech -r 0.20	1223 1228 1335	22.2% 22.6% 30.8%	15.9%
it-es_valid	webrtcvad -p 2	1075	14.4%	11.7%
es-en_test	webrtcvad -p 0	1116	11.4%	6.5%
es-en_valid	webrtcvad -p 0 webrtcvad -p 1	1082 1117	11.4% 14.6%	9.5%
pt-en_test	voxseg -s 0.90	1294	23.5%	15.1%
pt-en_valid	voxseg -s 0.95 inaspeech -r 0.05	1139 1199	11.7% 16.8%	16.5%

Table 5: Table displaying the best-found segmentation toolkit, the corresponding parameter, and the number of segments created. *The table also shows the percentage difference in segments counts and BLEU score compared to the scores given by the end-to-end model when utilizing the manual segmentation.

5.2 Automatic Segmentations Employing the Cascaded Model

Table 6 displays that by using the same automatic segmentation for the cascaded model, Voxseg only exceeds the other toolkits on three occasions by using a speech vs. non-speech threshold of -s 0.90, two of which segmentations were already best-performing in the conditions of an end-to-end system (i.e., Portuguese-Spanish and Portuguese-English on the test sets). While these segmentations remain best-performing, the BLEU score gap increased by approximately 1.5% in both tests.

While the previous scenario did not provide a single best-performing segmentation when employing the end-to-end model, inaSpeechSegmenter now has four experiments outperforming the other two toolkits on the cascaded model, primarily when Portuguese and

Italian are the source languages. In addition, the two inaSpeechSegmenter segmentations that previously led to equally accurate translations for the end-to-end multilingual translation system are now exceeding all other strategies.

Like Voxseg, the Webrtcvad toolkit leads to three best-performing translations, mainly by applying strategies leading to fewer segments that did well for both multilingual translation systems. (e.g., Spanish TEDx recordings). On the contrary, in the context of a cascaded system, the newly dominating inaSpeechSegmenter requires significantly more segments to outperform the other two toolkits (i.e., percentage difference varies between 16.8% to 22.6% compared to the man-made segmentations).

Language pair	Best segmentation	Segments count	Segments Difference*	BLEU score Difference*
pt-es_test	voxseg -s 0.90	1294	23.5%	19.3%
pt-es_valid	inaspeech -r 0.05	1199	16.8%	19.0%
it-en_test	inaspeech -r 0.15	1228	22.6%	13.1%
	inaspeech -r 0.20	1335	30.8%	
it-en_valid	webrtcvad -p 2	1075	14.4%	11.0%
it-es_test	inaspeech -r 0.15	1228	22.6%	16.6%
it-es_valid	voxseg -s 0.90	991	6.2%	12.8%
	webrtcvad -p 2	1075	14.4%	
es-en_test	webrtcvad -p 0	1116	11.4%	6.0%
es-en_valid	webrtcvad -p 1	1117	14.6%	7.7%
pt-en_test	voxseg -s 0.90	1294	23.5%	16.6%
pt-en_valid	inaspeech -r 0.05	1199	16.8%	17.4%

Table 6: Table displaying the best-found segmentation toolkit, the corresponding parameter, and the number of segments created. *The table also shows the percentage difference in segments counts and BLEU score compared to the scores given by the cascaded model when utilizing the manual segmentation.

5.3 All-around Automatic Segmentation Strategies

When putting all the results obtained with either of the two models altogether, it can be observed that just five out of the 17 applied segmentation strategies achieved best-performing tests in the vast majority of the cases (e.g., Voxseg -s 0.90 and Voxseg -s 0.95 on five and four tests, respectively).

It was earlier mentioned that, when using the manual segmentation, the Portuguese corpus leads to two opposite scenarios depending on the target language: The end-to-end model barely surpasses the cascaded approach when translating to Spanish (i.e., 0.3% on the test data), and the cascaded approach reaches better accuracies when translating to English. However, Figure 2 shows that when utilizing the segmentation produced by Voxseg, the end-to-end system maintains a more evident gap, especially in well-performing segmentations (i.e., percentage differences of 1.8% and 2.5% by using 1057 and 1294 segments, respectively). The figure reveals a clear cutoff point on all four tests when reaching 1294 segments, followed by a rapid and steady reduction in translation quality. Nevertheless, even in the case of over-segmentation of the TEDx recordings, the end-to-end model still surpasses the cascaded approach (e.g., 1.8% when using the highest speech vs. non-speech threshold creating 2161 segments).

Figure 2 also illustrates a different outcome for the Portuguese-English language pair: well-performing Voxseg segmentations led to the end-to-end model outperforming the cascaded approach. While the first four segmentation strategies applying -s thresholds ranging from 0.25 to 0.85 showed better accuracy on the cascaded model, higher thresholds led to the end-to-end surpassing the cascaded model in all remaining tests.

Figure 3 shows that the automatic segmentation performed by inaSpeechSegmenter does not lead to shifts in power in the end-to-end vs. cascaded challenge when considering the expected scenario (i.e., accuracies achieved by the manual segmentations). In contrast to the Voxseg segmentations in the previous figure, there are irregular predispositions in the gaps between the translation accuracies (e.g., the percentage gap on the Portuguese-Spanish pair increasing from 0.3% to 1.6% before decreasing abruptly back to 0.3% on the last strategy). The end-to-end model displays a cutoff point of around 1500 segments, BLEU scores on the Portuguese-Spanish language pair decreasing by 0.2 on the following strategy applying 100 more segments. On the other hand, the cascaded model keeps increasing after 1500 segments, achieving the best scores on the final segmentation strategy, leading to that swift drop in the gap between the Portuguese-Spanish language pair.

6 Discussion

The objects of this paper were to examine and compare modern VADs and discover the consequences of applying the resulting automatic segmentation for end-to-end and cascaded systems.

The paper discussed three modern yet, in many ways contrasting VADs, primarily when referring to the employed architecture, the extent of the training corpus, and the available modules of the frameworks (e.g., the possibility for gender recognition). However, despite these dissimilarities, there were also similar aspects among all investigated toolkits, such as the existence of adjustable inputs such as the speech vs. non-speech threshold. These adjustments play a vital role in the segmentation process and, therefore, the translation quality.

Segmentation strategy	BLEU scores and #segments counts on E2E and Cascaded models														
Test set	pt-es			pt-en			it-en			it-es			es-en		
	E2E	#seg	Casc.	E2E	#seg	Casc.	E2E	#seg	Casc.	E2E	#seg	Casc	E2E	#seg	Casc.
voxseg -s 0.90	32.9	1294	32.3	29.4	1294	29.3	25.9	998	25.0	29.5	998	27.7	37.8	1110	39.2
voxseg -s 0.95	30.9	1602	30.6	28.2	1602	28.0	26.7	1223	25.5	30.2	1223	27.9	36.6	1369	38.3
inaspeech -r 0.05	28.7	977	28.6	25.7	977	26.1	23.6	902	23.6	27.1	902	25.6	38.4	1288	39.5
inaspeech -r 0.15	31.6	1478	31.1	27.8	1478	28.3	26.2	1228	25.7	30.2	1228	28.1	36.8	1574	38.3
webrtcvad -p 2	31.7	1142	31.1	27.6	1142	27.7	25.6	932	24.5	29.0	932	26.8	37.3	1440	39.3

Table 7: Table displaying the BLEU scores and the corresponding number of segments on the Portuguese-Spanish, Portuguese-English, Italian-English, Italian-Spanish, and Spanish-English language pairs when employing five of the most dominating segmentation strategies and both the end-to-end and cascaded models.

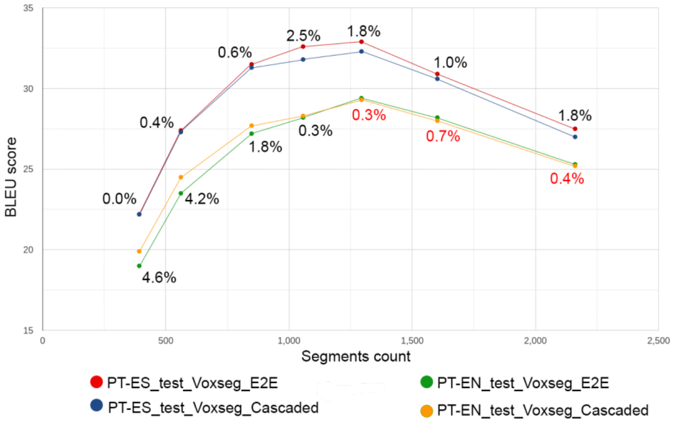

Figure 2: The figure shows the BLEU scores achieved using the Voxseg segmentation strategies on the Portuguese corpora when both models were employed for translation into Spanish and English. The figure also shows the percentage difference between the accuracies of the two models at each segmentation.

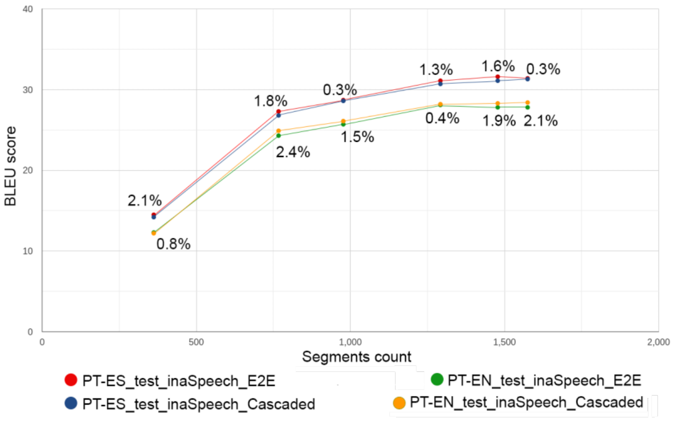

Figure 3: The figure shows the BLEU scores achieved using the inaSpeech segmentation strategies on the Portuguese corpora when both models were employed for translation into Spanish and English. The figure also shows the percentage difference between the accuracies of the two models at each segmentation.

The manual segmentation in the test and validation mTEDx dataset proved valuable in comparing and apprehending the importance of the results. The consequent translations hinted that, at least when adequate segmentations are employed, the end-to-end model should perform better translations or vice versa on the given source-target language pair.

The paper showed a dominant Voxseg VAD when employing the resulting segmentations to the end-to-end multilingual translation system. The previously deducted knowledge from the manual segmentations and Figure 2 help us understand the reasoning behind this prominence. The VAD consistently delivers robust results when employing reasonable segmentation strategies, generally reaching a stable cutoff point around 1200 segments followed by an inevitable decrease in quality when using too many segments.

However, the work also demonstrated that applying the same segmentations to the cascaded system leads to mixed outcomes in our tests, mainly affecting the reliability of our Voxseg segmentations. Despite inaSpeechSegmentations leading to inconsistent and notably inferior results for the end-to-end translation system, it is proving to be a more reasonable strategy for cascaded systems. In many circumstances, using more segments than the other two VADs, inaSpeechSegmenter can sustain an increasing trend in translation quality even though a cutoff point had already been reached when employing the same segmentation strategies for the end-to-end model. Therefore, employing over-segmentation through high thresholds has a significantly better consequence for inaSpeechSegmenter than the other two toolkits in the context of cascaded multilingual translation systems.

Lastly, the work conducted in this paper enforces the idea of a dominant new hybrid structure like VAD in the context of end-to-end translation systems. This statement was already proposed in the paper introducing the VAD (Wilkinson & Niesler, 2021). Later work in this field will tell us if similar VADs trained on more extensive corpora will dominate speech translation through both techniques or whether these structures are only exceeding in the case of end-to-end multilingual translation systems.

7 Conclusion

Research comparing modern VADs is scarce, especially on academic books. Similarly, the consequences when employing automatic segmentation to an end-to-end and a cascaded model is an understudied subject. This paper experiments by comparing three VADs using different architectures, analyzing the translation accuracy of these segmentations when employing an end-to-end and a cascaded translation system. The findings of this paper indicate that unique VADs using hybrid architectures such as Voxseg is a more suitable approach when using end-to-end models. The paper also shows that heavily trained VADs with denser structures, such as inaSpeechSegmenter, can perform better than the other reviewed toolkits when employing cascaded models, despite delivering worse results in the context of an end-to-end translation system.

8 Limitations

One limitation of the present study is the low number of tested parameters due to computation reasons. As a result, the accuracy of both the Voxseg and inaSpeechSegmenter VADs could have been slightly improved by optimizing the threshold values. Regardless, while these two parameters could potentially be optimized, the Webrtcvad VAD can only execute a maximum of four tests on each corpus because its parameter only accepts integers.

References

Berard, A., Besacier, L., Kocabiyikoglu, A., & Pietquin, O. (2018, 04). End-to-end automatic speech translation of audiobooks. In (p. 6224-6228). doi: 10.1109/ICASSP.2018.8461690

Bullington, K., & Fraser, J. M. (1959). Engineering aspects of tasi. *Transactions of the American Institute of Electrical Engineers, Part I: Communication and Electronics*, *78*(3), 256-260. doi: 10.1109/TCE.1959.6372993

Chaudhuri, S., Roth, J., Ellis, D. P. W., Gallagher, A. C., Kaver, L., Marvin, R., ... Xi, Z. (2018). Ava-speech: A densely labeled dataset of speech activity in movies. *CoRR*, *abs/1808.00606*. Retrieved from `http://arxiv.org/abs/1808.00606`

Doukhan, D., Lechapt, E., Evrard, M., & Carrive, J. (2018). Ina's mirex 2018 music and speech detection system. In *Music information retrieval evaluation exchange (mirex 2018).*

Giraudel, A., Carré, M., Mapelli, V., Kahn, J., Galibert, O., & Quintard, L. (2012, may). The repere corpus : a multimodal corpus for person recognition. In N. C. C. Chair) et al. (Eds.), *Proceedings of the eight international conference on language resources and evaluation (lrec'12).* Istanbul, Turkey: European Language Resources Association (ELRA).

Google. (2011). General overview of webrtc (web real-time communication).. Retrieved from `https://www.webrtc.org/reference/architecture`

Liu, D., & Niehues, J. (2021, August). Maastricht university's multilingual speech translation system for IWSLT 2021. In *Proceedings of the 18th international conference on spoken language translation (iwslt 2021)* (pp. 138–143). Bangkok, Thailand (online): Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2021.iwslt-1.15` doi: 10.18653/v1/2021.iwslt-1.15

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... Vesely, K. (2011, December). The kaldi speech recognition toolkit. In *Ieee 2011 workshop on automatic speech recognition and understanding.* IEEE Signal Processing Society. (IEEE Catalog No.: CFP11SRW-USB)

Rangarajan Sridhar, V. K., Chen, J., Bangalore, S., Ljolje, A., & Chengalvarayan, R. (2013, June). Segmentation strategies for streaming speech translation. In *Proceedings of the 2013 conference of the north American chapter of the association for computational linguistics: Human language technologies* (pp. 230–238). Atlanta, Georgia: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/N13-1023`

R. M. Bittner, E. J. H., & Bello, J. P. (2016). pysox: Leveraging the audio signal processing power of sox in python. In *Proceedings of the 17th international society for music information retrieval conference late breaking and demo papers.*

Salesky, E., Wiesner, M., Bremerman, J., Cattoni, R., Negri, M., Turchi, M., ... Post, M. (2021). *The multilingual tedx corpus for speech recognition and translation.*

Scheirer, E., & Slaney, M. (1997). Construction and evaluation of a robust multifeature speech/music discriminator. In *1997 ieee international conference on acoustics, speech, and signal processing* (Vol. 2, p. 1331-1334 vol.2). doi: 10.1109/ICASSP.1997.596192

Snyder, D., Chen, G., & Povey, D. (2015). *Musan: A music, speech, and noise corpus.*

Sperber, M., & Paulik, M. (2020). *Speech translation and the end-to-end promise: Taking stock of where we are.*

Sturm, B. L. (2014, Apr). The state of the art ten years after a state of the art: Future research in music information retrieval. *Journal of New Music Research*, *43*(2), 147–172. Retrieved from `http://dx.doi.org/10.1080/09298215.2014.894533` doi: 10.1080/09298215.2014.894533

Wilkinson, N., & Niesler, T. (2021). A hybrid CNN-BiLSTM voice activity detector. In *Proc. ieee international conference on acoustics, speech, and signal processing (icassp).* Toronto, Canada.

Wiseman, J. (2020). Python interface to the webrtc voice activity detector. In *Python interface to the webrtc voice activity detector.*

A Appendix

Segmentation type	Segmentation strategy	E2E (E4)			Cascaded (E4+M1)		
		BLEU	#segments count	#words per hypothesized line (mean ± SD)	BLEU	#segments count	#words per hypothesized line (mean ± SD)
es-en_test_ voxseg	-s 0.25	27.8	465	24.93 ± 19.34	29.6	465	25.96 ± 19.31
	-s 0.50	33.2	634	20.93 ± 15.01	34.0	634	21.11 ± 15.33
	-s 0.75	36.8	876	16.75 ± 12.60	38.1	876	16.95 ± 13.11
	-s 0.85	37.5	1002	15.12 ± 11.55	39.1	1002	15.30 ± 12.03
	-s 0.90	37.8	1110	13.88 ± 10.71	39.2	1110	14.0 ± 11.04
	-s 0.95	36.6	1369	11.62 ± 9.08	38.3	1369	11.62 ± 9.08
	-s 0.99	33.3	1889	8.66 ± 6.53	34.9	1889	8.72 ± 6.61
es-en_test_ inaspeech	-r 0.01	33.3	779	15.87 ± 14.52	34.9	779	16.08 ± 15.14
	-r 0.03	37.2	1095	13.39 ± 11.30	39.0	1095	13.46 ± 11.47
	-r 0.05	38.4	1288	11.98 ± 9.86	39.5	1288	12.10 ± 10.10
	-r 0.10	37.3	1477	10.55 ± 8.42	38.8	1477	10.69 ± 8.72
	-r 0.15	36.8	1574	9.98 ± 7.87	38.3	1574	10.07 ± 8.02
	-r 0.20	36.6	1624	9.71 ± 7.65	38.1	1624	9.77 ± 7.93
es-en_test_ webrtcvad	-p 0	38.7	1116	13.84 ± 10.76	40.5	1116	14.00 ± 11.33
	-p 1	38.3	1191	13.02 ± 10.12	40.2	1191	13.19 ± 10.61
	-p 2	37.3	1440	11.06 ± 8.51	39.3	1440	11.12 ± 8.68
	-p 3	36.0	1736	8.42 ± 6.18	36.0	1736	8.44 ± 6.23

Table 8: BLEU scores at the talk level for the Spanish-English reference text for three segmentation VADs using the E2E and Cascaded approaches.

Segmentation type	Segmentation strategy	E2E (E4)			Cascaded (E4+M1)		
		BLEU	#segments count	#words per hypothesized line (mean ± SD)	BLEU	#segments count	#words per hypothesized line (mean ± SD)
pt-en_test_ voxseg	-s 0.25	19.0	392	24.37 ± 21.16	19.9	392	24.75 ± 21.00
	-s 0.50	23.5	561	19.83 ± 17.09	24.5	561	20.17 ± 17.50
	-s 0.75	27.2	848	15.16 ± 13.49	27.7	848	15.38 ± 14.20
	-s 0.85	28.2	1057	12.88 ± 11.41	28.3	1057	12.98 ± 11.92
	-s 0.90	29.4	1294	11.43 ± 10.55	29.3	1294	11.41 ± 10.69
	-s 0.95	28.2	1602	9.60 ± 8.70	28.0	1602	9.62 ± 8.83
	-s 0.99	25.3	2161	6.89 ± 5.97	25.2	2161	6.89 ± 6.05
pt-en_test_ inaspeech	-r 0.01	12.3	362	20.63 ± 26.25	12.2	362	20.77 ± 23.84
	-r 0.03	24.3	767	14.83 ± 17.08	24.9	767	15.04 ± 17.01
	-r 0.05	25.7	977	12.54 ± 13.87	26.1	977	12.68 ± 14.16
	-r 0.10	28.1	1292	10.53 ± 10.49	28.2	1292	10.51 ± 10.54
	-r 0.15	27.8	1478	9.51 ± 9.21	28.3	1478	9.57 ± 9.22
	-r 0.20	27.8	1575	8.99 ± 8.41	28.4	1575	9.04 ± 8.53
pt-en_test_ webrtcvad	-p 0	25.5	861	13.74 ± 12.16	25.8	861	13.93 ± 12.89
	-p 1	26.4	930	13.09 ± 11.65	26.2	930	13.14 ± 11.98
	-p 2	27.6	1142	11.46 ± 10.87	27.7	1142	11.49± 11.04
	-p 3	27.0	1990	7.96 ± 6.48	27.7	1990	7.93 ± 6.54

Table 9: BLEU scores at the talk level for the Portuguese-English reference text for three segmentation VADs using the E2E and Cascaded approaches.

Segmentation type	Segmentation strategy	E2E (E4)			Cascaded (E4+M1)		
		BLEU	#segments count	#words per hypothesized line (mean ± SD)	BLEU	#segments count	#words per hypothesized line (mean ± SD)
pt-es_test_voxseg	-s 0.25	22.2	392	25.75± 21.52	22.2	392	25.79± 22.54
	-s 0.50	27.4	561	21.07± 18.23	27.3	561	21.03±18.25
	-s 0.75	31.5	848	15.86± 14.40	31.3	848	16.00± 14.61
	-s 0.85	32.6	1057	13.45± 12.32	31.8	1057	13.42± 12.25
	-s 0.90	32.9	1294	11.77± 11.06	32.3	1294	11.76± 10.95
	-s 0.95	30.9	1602	9.86± 9.33	30.6	1602	9.90± 9.17
	-s 0.99	27.5	2161	6.94± 6.31	27.0	2161	6.95± 6.27
pt-es_test_inaspeech	-r 0.01	14.5	362	21.63± 28.34	14.2	362	21.82± 27.45
	-r 0.03	27.3	767	15.41± 17.94	26.8	767	15.34± 17.31
	-r 0.05	28.7	977	12.91± 14.94	28.6	977	13.07± 15.41
	-r 0.10	31.1	1292	10.71± 11.08	30.7	1292	10.68± 10.86
	-r 0.15	31.6	1478	9.72± 9.78	31.1	1478	9.74± 9.67
	-r 0.20	31.4	1575	9.17± 8.94	31.3	1575	9.18± 8.87
pt-es_test_webrtcvad	-p 0	29.4	861	14.26± 13.32	29.0	861	14.3± 13.48
	-p 1	30.2	930	13.57± 12.50	29.7	930	13.48± 12.45
	-p 2	31.7	1142	11.78± 11.83	31.1	1142	11.68± 11.14
	-p 3	30.2	1990	8.02± 6.83	30.1	1990	8.06± 6.86

Table 10: BLEU scores at the talk level for the Portuguese-Spanish reference text for three segmentation VADs using the E2E and Cascaded approaches.

Segmentation type	Segmentation strategy	E2E (E4)			Cascaded (E4+M1)		
		BLEU	#segments count	#words per hypothesized line (mean ± SD)	BLEU	#segments count	#ords per hypothesized line (mean ± SD)
it-en_test_voxseg	-s 0.25	15.0	375	26.81 ± 24.64	14.7	375	27.47 ± 25.91
	-s 0.50	19.4	512	22.93 ± 21.10	18.8	512	23.31 ± 21.58
	-s 0.75	22.4	699	19.00 ± 18.21	21.9	699	19.23 ± 17.80
	-s 0.85	24.5	852	16.56 ± 15.77	23.8	852	17.33 ± 16.23
	-s 0.90	25.9	998	15.54 ± 15.00	25.0	998	15.81 ± 14.83
	-s 0.95	26.7	1223	13.71 ± 12.79	25.5	1223	13.94 ± 13.31
	-s 0.99	25.5	1743	10.26 ± 9.30	23.9	1743	10.32 ± 9.56
it-en_test_inaspeech	-r 0.01	16.7	475	22.27 ± 22.79	16.9	475	22.90 ± 23.17
	-r 0.03	22.3	772	17.23 ± 17.00	21.8	772	17.36 ± 17.36
	-r 0.05	23.6	902	15.57 ± 15.47	23.6	902	16.14 ± 16.63
	-r 0.10	25.7	1097	14.04 ± 13.71	25.5	1097	14.39 ± 14.63
	-r 0.15	26.2	1228	13.04 ± 12.36	25.7	1228	13.37 ± 13.47
	-r 0.20	26.6	1335	12.43 ± 11.60	25.7	1335	12.70 ± 12.38
it-en_test_webrtcvad	-p 0	22.2	644	19.86 ± 17.71	20.9	644	19.96 ± 17.63
	-p 1	23.0	707	18.81 ± 17.00	22.1	707	19.14 ± 18.16
	-p 2	25.6	932	16.08 ± 14.63	24.5	932	16.06 ± 14.72
	-p 3	25.9	1652	10.97 ± 9.22	24.8	1652	11.00 ± 9.31

Table 11: BLEU scores at the talk level for the Italian-English reference text for three segmentation VADs using the E2E and Cascaded approaches.

Segmentation type	Segmentation strategy	E2E (E4)		Cascaded (E4+M1)	
		BLEU	#segments count	BLEU	#segments count
it-es_test_ voxseg	-s 0.25	17.2	375	16.5	375
	-s 0.50	22.3	512	20.8	512
	-s 0.75	25.7	699	23.8	699
	-s 0.85	27.7	852	25.7	852
	-s 0.90	29.5	998	27.7	998
	-s 0.95	30.2	1223	27.9	1223
	-s 0.99	28.2	1743	25.9	1743
it-es_test_ inaspeech	-r 0.01	19.6	475	17.9	475
	-r 0.03	25.3	772	23.5	772
	-r 0.05	27.1	902	25.6	902
	-r 0.10	29.3	1097	27.8	1097
	-r 0.15	30.2	1228	28.1	1228
	-r 0.20	30.2	1335	27.6	1335
it-es_test_ webrtcvad	-p 0	24.5	644	22.7	644
	-p 1	26.0	707	24.0	707
	-p 2	29.0	932	26.8	932
	-p 3	29.1	1652	27.1	1652

Table 12: BLEU scores at the talk level for the Italian-Spanish reference text for three segmentation VADs using the E2E and Cascaded approaches.

Segmentation type	Segmentation strategy	IT-EN_valid				PT-ES_valid				ES-EN_valid			
		E2E (E4)		Cascaded (E4+M1)		E2E (E4)		Cascaded (E4+M1)		E2E (E4)		Cascaded (E4+M1)	
		BLEU	#segments count	BLEU	#segments count	BLEU	#segments count	BLEU	#segments count	BLEU	#segments count	BLEU	#segments count
voxseg	-s 0.25	17.9	432	16.1	432	20.3	394	15.8	394	27.0	530	28.5	530
	-s 0.50	20.0	555	19.1	555	27.1	511	24.7	511	30.0	687	31.6	687
	-s 0.75	22.9	744	22.0	744	31.2	679	30.2	679	32.3	850	34.3	850
	-s 0.85	23.5	880	23.1	880	34.0	786	33.0	786	33.7	972	34.9	972
	-s 0.90	24.4	991	23.4	991	35.3	890	34.5	890	34.0	1071	35.0	1071
	-s 0.95	24.5	1190	23.5	1190	37.1	1139	35.7	1139	32.8	1291	33.9	1291
	-s 0.99	23.2	1655	22.4	1655	34.3	1659	33.4	1659	29.8	1784	30.8	1784
inaspeech	-r 0.01	18.3	523	14.7	523	28.7	451	27.1	451	31.6	853	33.2	853
	-r 0.03	23.8	853	22.5	853	35.7	1004	35.3	1004	34.1	1245	35.8	1245
	-r 0.05	24.5	977	23.6	977	37.0	1199	36.0	1199	34.0	1344	35.5	1344
	-r 0.10	24.9	1176	23.8	1176	36.0	1424	35.0	1424	33.2	1486	34.5	1486
	-r 0.15	24.7	1287	23.5	1287	35.3	1536	34.5	1536	32.6	1561	34.3	1561
	-r 0.20	24.0	1372	23.4	1372	35.2	1608	34.2	1608	32.5	3204	33.9	3204
webrtcvad	-p 0	23.6	782	23.4	782	33.0	883	32.4	883	34.2	1082	35.5	1082
	-p 1	24.3	840	23.0	840	33.8	936	33.1	936	34.2	1117	36.0	1117
	-p 2	25.3	1075	24.0	1075	35.7	1171	35.1	1171	32.7	1537	35.0	1537
	-p 3	24.2	1673	22.9	1673	33.6	1844	33.1	1844	30.5	1848	32.1	1848

Table 13: BLEU scores at the talk level for the Italian-Spanish, Portuguese-Spanish and Spanish-English validation reference texts for three segmentation VADs using the E2E and Cascaded approaches.