MuST-C: a Multilingual Speech Translation Corpus

Mattia Antonino Di Gangi[1,2], **Roldano Cattoni**[1], **Luisa Bentivogli**[1],
Matteo Negri[1] and **Marco Turchi**[1]
[1]Fondazione Bruno Kessler
[2]University of Trento
Trento, Italy
{digangi,cattoni,bentivo,negri,turchi}@fbk.eu

Abstract

Current research on spoken language translation (SLT) has to confront with the scarcity of sizeable and publicly available training corpora. This problem hinders the adoption of neural end-to-end approaches, which represent the state of the art in the two parent tasks of SLT: automatic speech recognition and machine translation. To fill this gap, we created MuST-C, a multilingual speech translation corpus whose size and quality will facilitate the training of end-to-end systems for SLT from English into 8 languages. For each target language, MuST-C comprises at least 385 hours of audio recordings from English TED Talks, which are automatically aligned at the sentence level with their manual transcriptions and translations. Together with a description of the corpus creation methodology (scalable to add new data and cover new languages), we provide an empirical verification of its quality and SLT results computed with strong baseline system on each language direction.

1 Introduction

Besides the increased computing power, the recent surge of neural end-to-end approaches to natural language processing tasks has been stoked by the increased availability of data. For instance, when supported by sizeable training corpora, the robustness and the strong generalization capabilities of neural networks led to their dominance over previous paradigms both in automatic speech recognition (ASR (Chiu et al., 2018)) and machine translation (MT (Bojar et al., 2018)).

Compared to its two parent research areas, spoken language translation (SLT) has not shown such a steady progress yet. Despite recent claims by big industry players about the effectiveness of end-to-end learning (Weiss et al., 2017; Jia et al., 2018), its adoption does not yet represent the mainstream solution to the SLT task. One of the main obstacles to a stable dominance of the end-to-end paradigm also in this area is the scarcity of training corpora. While cascade ASR+MT solutions can exploit the wealth of task-specific data available for each of the two tasks,[1] the situation for end-to-end model training is much less favourable. As shown in Table 1, few publicly available corpora exist, their language coverage is rather limited and, most importantly, their size is often too small (less than 100 hours of translated audio) for training data-hungry neural models.[2]

Corpus	Languages	Hours
Niehues et al. (2018)	En→De	273
Kocabiyikoglu et al. (2018)	En→Fr	236
Tohyama et al. (2005)	En↔Jp	182
Paulik and Waibel (2009)	En→Es	111
	Es→En	105
Post et al. (2013)	En→Es	38
Stüker et al. (2012)	De→En	37
Shimizu et al. (2014)	En↔Jp	22
Federmann and Lewis (2017)	En↔Jp/Zh	22
Bendazzoli and Sandrelli (2005)	En↔It/Es	18
	It↔Es	
Bérard et al. (2016)	Fr→En	17
Federmann and Lewis (2016)	En↔Fr/De	8
Woldeyohannis et al. (2017)	Am→En	7
Godard et al. (2017)	Mboshi→Fr	4

Table 1: Publicly available SLT corpora. The two most recent resources (also known as *IWSLT18* and *Augmented LibriSpeech*) are also the largest ones. Though considerably smaller, the *Fisher and Callhome* corpus described in (Post et al., 2013) is among the most widely used ones in previous research.

To circumvent the problem, neural SLT approaches currently rely on: *i)* large proprietary corpora (Jia et al., 2018), *ii)* multitask learning

[1]In resource-rich conditions, ASR and MT training often builds on thousands of hours of transcribed speech and tens of millions of parallel sentences, respectively.

[2]Besides the corpora reported in Table 1, several smaller ($<$ 4 hours) freely-available datasets have been created (*e.g.* the IWSLT evaluation campaign development and test sets from 2010 to 2017 and the Griko-Italian corpus by Boito et al. (2018)).

Proceedings of NAACL-HLT 2019, pages 2012–2017
Minneapolis, Minnesota, June 2 - June 7, 2019.

(Weiss et al., 2017; Anastasopoulos and Chiang, 2018; Bérard et al., 2018), *iii)* encoder/decoder pre-training (Bansal et al., 2018; Bérard et al., 2018), *iv)* synthesized speech data (Bérard et al., 2016), or *v)* machine-translated target text data (Bérard et al., 2018). Though effective, solutions *ii)* and *iii)* assume the availability of ASR and MT data, which is not always guaranteed (especially in low-resource language settings). Solutions *iv)* and *v)*, instead, rely on training material derived from sub-optimal automatic data creation/augmentation procedures. This situation calls for initiatives towards the creation of large, high-quality multilingual corpora suitable to explore end-to-end SLT in more favorable conditions similar to condition *i)*. Along this direction, our contributions are:

- A large (∼400 hours of speech per language) multilingual corpus for SLT from English into 8 languages (German, Spanish, French, Italian, Dutch, Portuguese, Romanian and Russian);
- An empirical verification of its quality;
- ASR, MT and SLT results computed with strong baseline systems on each language direction.

MuST-C is released under a Creative Commons license, Attribution - Non Commercial - No Derivatives (CC BY NC ND 4.0 International), and is freely downloadable at `mustc.fbk.eu`

2 Corpus Creation Methodology

Must-C was created pursuing high quality as well as large size, speaker variety (male/female, native/non-native) and coverage in terms of topics and languages. To achieve these objectives, similar to (Niehues et al., 2018), we started from English TED Talks, in which a variety of speakers discuss topics spanning from business to science and entertainment. Most importantly, the fact that TED talks are often manually transcribed and translated sets ideal conditions for creating an SLT corpus from high-quality text material. Although the initial data are similar to those used to build the IWSLT18 corpus, our methodology is different. Inspired by Kocabiyikoglu et al. (2018), it exploits automatic alignment procedures, first at the text level (between transcriptions and translations) and then with the corresponding audio segments.

More in detail, for each target language L_i, the (English-L_i) section of MuST-C is created as follows. First, for all the English talks available from the TED website,[3] we download the videos and the HTML files containing the manual transcriptions and their translation into L_i.[4]

Then, the plain text transcription and the translation of each talk are split at the sentence level based on strong punctuation marks and aligned using the Gargantua sentence alignment tool (Braune and Fraser, 2010). This step produces a bilingual text corpus aligned at the sentence level.

In the third step, the English side of this bilingual corpus is aligned to the corresponding audio track extracted from the video. This is done using Gentle,[5] an off-the-shelf English forced-aligner built on the Kaldi ASR toolkit (Povey et al., 2011).

Next, the audio-text alignments are processed to create a YAML file containing time information (*i.e.* start and duration) for each sentence. In this processing step, two filters are applied to weed out potentially noisy segments, or entire talks, based on the number of words that were not aligned by Gentle. First, entire talks are discarded if the proportion of unrecognized words is equal or greater than 15% of the total. This threshold was determined after a manual analysis of 73 talks (those with the highest percentage of unrecognized words). The analysis showed that these cases are representative of different types of noise like: *i)* non-English speech, *ii)* long silences, *iii)* music, non-transcribed songs and videos played during the talk, and *iv)* wrong transcriptions (*e.g.* captions from other talks in the material downloaded from the TED website). The second rule applies to the single sentences of the talks that passed the first filter, and removes those in which none of the words was aligned by Gentle.[6]

In the last step, the log Mel 40-dimensional filter-bank features – commonly used as input representation for ASR (Graves et al., 2013) and SLT (Weiss et al., 2017) – are extracted from the

[3] `www.ted.com` – dump of April 2018.

[4] All talks have manual captions, which were also translated into many languages by volunteers. The language coverage of the translations depends on several factors like the age of the talk (the old ones often have more translations), the popularity of its topic and the availability of volunteer translators for a given language.

[5] `github.com/lowerquality/gentle`

[6] The effectiveness of this filtering criterion was manually verified on random samples. More aggressive solutions will be explored for future releases of the corpus.

Tgt	#Talk	#Sent	Hours	src w	tgt w
De	2,093	234K	408	4.3M	4.0M
Es	2,564	270K	504	5.3M	5.1M
Fr	2,510	280K	492	5.2M	5.4M
It	2,374	258K	465	4.9M	4.6M
Nl	2,267	253K	442	4.7M	4.3M
Pt	2,050	211K	385	4.0M	3.8M
Ro	2,216	240K	432	4.6M	4.3M
Ru	2,498	270K	489	5.1M	4.3M

Table 2: Statistics for each section of MuST-C.

aligned audio using the XNMT tool (Neubig et al., 2018).[7]

Table 2 provides basic statistics for the 8 sections of the MuST-C corpus. Comparing the 4^{th} column with the numbers reported in Table 1, it is worth noting that, in terms of hours of transcribed/translated speech, each section is larger than any existing publicly available SLT resource.

3 Experiments

In this section we present two sets of experiments, which are respectively aimed to: *i)* empirically assess the quality of the MuST-C corpus (Section 3.3) and *ii)* compute baseline ASR, MT, and SLT results for future comparisons (Section 3.4).

In these experiments, the *audio-transcription* alignments of MuST-C are used to train and evaluate ASR models, *transcription-translation* alignments are used for the MT models, and *audio-translation* alignments are used for the SLT models.

3.1 ASR, MT and SLT Models

ASR and SLT. For our experiments in ASR and SLT we use the same neural architecture. This setting allows us to use the encoder of the ASR models to initialize the weights of the SLT encoders and achieve a faster convergence (Bansal et al., 2018). Our SLT architecture is a variant of the system proposed by Bérard et al. (2018), which we re-implemented in the fairseq toolkit (Gehring et al., 2017). The system relies on an attentional encoder-decoder model that takes in input sequences of audio features and outputs the target sequence at the character level. The encoder processes the input with two consecutive fully-connected layers to expand the size of the representation, followed by two 2D strided convolutional layers that reduce the sequence length. The output of the convolutions is then processed by three stacked LSTMs (Hochreiter and Schmidhuber, 1997). The decoder consists of a two-layered deep transition (Pascanu et al., 2014) LSTM with an attention network based on the general soft attention score (Luong et al., 2015). The final output of the decoder is a function of the concatenation of the LSTM output, the context vector and the previous-character embedding.

MT. For the MT experiments we use the open source version of ModernMT.[8] The system is based on the Transformer (Vaswani et al., 2017) architecture, which represents the state of the art in NMT (Bojar et al., 2018). The encoder consists of a stack of 6 layers, each containing a sequence of two sub-layers, a self-attention network based on multi-head attention, and a position-wise feed-forward layer. The decoder layers have an additional sub-layer: between the self attention and the position-wise feed-forward layer they have an encoder-decoder multi-head attention. All the sub-layers in both the encoder and decoder are preceded by layer normalization and are followed by residual connections.

3.2 Data Processing and Evaluation Metrics

In our experiments, texts are tokenized and punctuation is normalized. Furthermore, the English texts are lowercased, while the target language texts are split into characters still preserving the word boundaries. For MT, we segment the English words with the BPE algorithm (Sennrich et al., 2015) using a maximum of $30K$ merge operations. The output generation of all models is performed using beam search with a beam size of 5.

ASR performance is measured with word error rate (WER) computed on lower-cased, tokenized texts without punctuation. MT and SLT results are computed with BLEU (Papineni et al., 2002).

3.3 Experiment 1: Corpus Quality

As observed in Section 2, each section of MuST-C is larger than any other existing publicly available SLT corpus. The usefulness of a resource, however, is not only a matter of size but also of quality (in this case, the quality of the *audio-transcription-translation* alignments). For an empirical verification of this aspect, we experimented with two comparable datasets. One is

[7] github.com/neulab/xnmt

[8] www.modernmt.eu

the TED-derived English-German IWSLT18 corpus (Niehues et al., 2018), which is built following a pipeline that performs segment extraction and alignment based on time information (*i.e.* start and end position of each segment in the SubRip Text (SRT) files) instead of text-level alignments. The other is the English-German subset of MuST-C derived from the same TED Talks used to build the IWSLT18 corpus. On one side (MuST-C), the number of segments, their length, and the overall corpus quality depend on text-level alignments. On the other side (IWSLT18), they depend on matching time stamps. This strategy, however, has some drawbacks. First, as pointed out by (Niehues et al., 2018; Liu et al., 2018; Di Gangi et al., 2018), the use of time information brings some noise in the corpus. Second, it often results in utterance-level alignment (based on speakers' pauses in the original audio). Compared to sentence-level alignment, this level of granularity can be sub-optimal during model training (*e.g.* for MT and SLT, learning from complete sentences is easier than learning from phrases). Finally, time information about the recorded speech is not always available: bypassing this need would make the method replicable on other data (not only TED-like).

Though initialized with the same set of $1,619$ talks, the two pipelines produce different corpora. As shown in Table 3, our approach filters out 58 entire talks ($\sim$3.6% of the total) but the final number of segments, their corresponding audio duration and their average length (in words) are larger.

Corpus	#Talk	#Sent	Hours	src w	tgt w
IWSLT18	1,619	176K	280	2.7M	2.5M
MuST-C	1,561	179K	313	3.3M	3.1M

Table 3: Statistics of the English-German corpora created by applying the IWSLT18 and MuST-C pipelines to the same initial set of $1,619$ TED Talks.

Each corpus was divided into training, development and test. Development and test contain segments from randomly selected common talks (*i.e.* those preserved by the MuST-C pipeline). Their size is respectively $2.3K$ (from 28 talks) and $2.1K$ segments (from 26 talks). The test portions were concatenated to create a balanced test set ($4.2K$ segments) containing half of the instances from the IWSLT18 corpus and half from MuST-C. The remaining material was used to separately train ASR, MT and SLT models on homogeneous data from either of the two corpora (*i.e.* three systems per corpus). All the systems are evaluated on the common test set.

Training set	ASR (↓)	MT (↑)	SLT (↑)
IWSLT18	42.15	24.90	8.94
MuST-C	32.05	25.46	12.25

Table 4: Performance of ASR, MT and SLT systems trained with En-De IWSLT18 and MuST-C data.

Tgt	ASR (↓)	MT (↑)	SLT (↑)
De	27.00	28.09	12.93
Es	26.61	34.16	18.20
Fr	25.81	42.23	22.29
It	26.38	30.40	14.95
Nl	26.55	33.43	18.20
Pt	28.00	32.44	17.10
Ro	27.61	28.16	13.35
Ru	26.97	18.30	7.22

Table 5: Baseline ASR, MT and SLT results for each language direction.

Table 4 shows that the models trained on MuST-C data achieve better results on the balanced test set in all the three tasks. In particular: *i)* a reduction of 10.1 WER points in ASR indicates a higher quality of *audio-transcription* alignments, *ii)* a BLEU increase of 0.56 points in MT indicates a similar quality for *transcription-translation* alignments, and *iii)* a BLEU increase of 3.31 points in SLT indicates a higher quality of *audio-translation* alignments. We consider these results as evidence of the reliability of our corpus creation methodology. Being the same for all the language pairs, we expect this procedure to end up in comparable quality for all the 8 sections of MuST-C.

3.4 Experiment 2: Baseline Results

We finally present baseline results computed, for all the three tasks, on each section of MuST-C. Also for these experiments, development and test data are created with segments from talks that are common to all the languages. Their size is respectively $1.4K$ (from 11 talks) and $2.5K$ segments (from 27 talks). The remaining data (of variable size depending on the language pairs) are used for training. For the sake of replicability, these splits are preserved in the released version of MuST-C.

The results in Table 5 lead to the following observations. First, though not directly comparable since they are computed on different test sets, English-German results are in line (actually

higher, since they are produced by models built on larger training data) with those presented in Section 3.3. This indicates that the level of quality observed in the previous experiments with a subset of the training data is preserved by the whole material released for this language pair. Second, looking at the other language pairs, ASR, MT and SLT results are comparable with the English-German scores. Besides normal fluctuations in the optimization of the neural models, performance differences are coherent with: *i)* the relative difficulty of each target language (*e.g.* Russian is more difficult due to high inflection) and *ii)* the variable quantity of training data available (*e.g.* French has the largest training set, see Table 2). Overall, these explainable differences suggest that our corpus creation methodology yields homogeneous quality for all the languages covered by MuST-C.

4 Conclusion and Future Work

We presented MuST-C, a *Mu*ltilingual *S*peech *T*ranslation *C*orpus built to address the need of resources for training data-hungry neural SLT models. To the best of our knowledge, to date MuST-C is the largest publicly available corpus of this kind. In its current version, it comprises the English transcription and the translations into 8 target languages of at least 385 hours of speech (up to 504) per language. Thanks to a scalable corpus creation procedure initialized with constantly expanding TED talks data, future extensions will increase the coverage of the already present target languages and introduce new ones.

MuST-C is released under a Creative Commons license, Attribution - Non Commercial - No Derivatives (CC BY NC ND 4.0 International), and is freely downloadable at `mustc.fbk.eu`

Acknowledgments

The authors gratefully acknowledge NVIDIA Corporation for the donation of the Tesla K80 and GeForce GTX 1080 Ti GPUs used for this research.

References

Antonios Anastasopoulos and David Chiang. 2018. Tied Multitask Learning for Neural Speech Translation. In *Proceedings of NAACL-HLT 2018*, pages 82–91, New Orleans, Louisiana.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2018. Pre-Training on High-Resource Speech Recognition Improves Low-Resource Speech-to-Text Translation. *arXiv preprint arXiv:1809.01431*.

Claudio Bendazzoli and Annalisa Sandrelli. 2005. An Approach to Corpus-based Interpreting Studies: Developing EPIC (European Parliament Interpreting Corpus). In *Proceedings of the EU-High-Level Scientific Conference Series MuTra 2005Challenges of Multidimensional Translation*, pages 149–160, Saarbrücken, Germany.

Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. 2018. End-to-End Automatic Speech Translation of Audiobooks. In *Proceedings of ICASSP 2018*, pages 6224–6228, Calgary, Alberta, Canada.

Alexandre Bérard, Olivier Pietquin, Laurent Besacier, and Christophe Servan. 2016. Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. In *Proceedings of the NIPS Workshop on end-to-end learning for speech and audio processing*, Barcelona, Spain.

Marcely Zanon Boito, Antonios Anastasopoulos, Marika Lekakou, Aline Villavicencio, and Laurent Besacier. 2018. A Small Griko-Italian Speech Translation Corpus. *CoRR*, abs/1807.10740.

Ondrej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 Conference on Machine Translation (WMT18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 272–307, Belgium, Brussels.

Fabienne Braune and Alexander Fraser. 2010. Improved Unsupervised Sentence Alignment for Symmetrical and Asymmetrical Parallel Corpora. In *Proceedings of COLING 2010*, pages 81–89, Beijing, China.

Chung-Cheng Chiu, Tara Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, et al. 2018. State-of-the-art Speech Recognition With Sequence-to-Sequence Models. In *Proceedings of ICASSP 2018*, pages 4774–4778, Calgary, Alberta, Canada.

Mattia Antonino Di Gangi, Roberto Dessì, Roldano Cattoni, Matteo Negri, and Marco Turchi. 2018. Fine-tuning on Clean Data for End-to-End Speech Translation: FBK@ IWSLT 2018. In *Proceedings of IWSLT 2018*, Bruges, Belgium.

Christian Federmann and William D Lewis. 2016. Microsoft Speech Language Translation (MSLT) Corpus: The IWSLT 2016 Release for English, French and German. In *Proceedings of IWSLT 2016*, Seattle, USA.

Christian Federmann and William D Lewis. 2017. The Microsoft Speech Language Translation (MSLT) Corpus for Chinese and Japanese: Conversational Test data for Machine Translation and Speech

Recognition. In *Proceedings of the 16th Machine Translation Summit*, Nagoya, Japan.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional Sequence to Sequence Learning. In *Proceedings of ICML 2017*, pages 1243–1252, Sydney, Australia.

Pierre Godard, Gilles Adda, Martine Adda-Decker, Juan Benjumea, Laurent Besacier, et al. 2017. A Very Low Resource Language Speech Corpus for Computational Language Documentation Experiments. *arXiv preprint arXiv:1710.03501*.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey E. Hinton. 2013. Speech Recognition with Deep Recurrent Neural Networks. In *Proceedings of ICASSP 2018*, pages 6645–6649, Vancouver, BC, Canada.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural computation*, 9(8).

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron-J. Weiss, Yuan Cao, Chung-Cheng Chiu, Stella-Laurenzo Ari, and Yonghui Wu. 2018. Leveraging Weakly Supervised Data to Improve End-to-End Speech-to-Text Translation. *ArXiv e-prints arXiv:1811.02050*.

Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. 2018. Augmenting Librispeech with French Translations: A Multimodal Corpus for Direct Speech Translation Evaluation. In *Proceedings of LREC 2018*, Miyazaki, Japan.

Dan Liu, Junhua Liu, Wu Guo, Shifu Xiong, Zhiqiang Ma, Rui Song, Chongliang Wu, and Quan Liu. 2018. The USTC-NEL Speech Translation system at IWSLT 2018. In *Proceedings of IWSLT 2018*, Bruges, Belgium.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of EMNLP 2015*, pages 1412–1421, Lisbon, Portugal.

Graham Neubig, Matthias Sperber, Xinyi Wang, Matthieu Felix, Austin Matthews, et al. 2018. XNMT: The eXtensible Neural Machine Translation Toolkit. In *Proceedings of AMTA 2018*, pages 185–192, Boston, MA.

Jan Niehues, Roldano Cattoni, Sebastian Stüker, Mauro Cettolo, Marco Turchi, and Marcello Federico. 2018. The IWSLT 2018 Evaluation Campaign. In *Proceedings of IWSLT 2018*, Bruges, Belgium.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL 2002*, Philadelphia, PA, USA.

Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. How to Construct Deep Recurrent Neural Networks. In *Proceedings of ICLR 2014*, Banff, Canada.

Matthias Paulik and Alex Waibel. 2009. Automatic Translation from Parallel Speech: Simultaneous Interpretation as MT Training Data. In *Proceedings of ASRU 2009*, pages 496–501, Merano, Italy.

Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. 2013. Improved Speech-to-Text Translation with the Fisher and Callhome Spanish–English Speech Translation Corpus. In *Proceedings of IWSLT 2013*, Heidelberg, Germany.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, K. Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. 2011. The Kaldi Speech Recognition Toolkit. In *Proceedings of ASRU 2011*, pages 1–4, Big Island, Hawaii, USA.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural Machine Translation of Rare Words with Subword Units. *arXiv preprint arXiv:1508.07909*.

Hiroaki Shimizu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Collection of a Simultaneous Translation Corpus for Comparative Analysis. In *Proceedings of LREC 2014*, Reykjavik, Iceland.

Sebastian Stüker, Florian Kraft, Christian Mohr, Teresa Herrmann, Eunah Cho, and Alex Waibel. 2012. The KIT Lecture Corpus for Speech Translation. In *Proceedings of LREC-2012*, Istanbul, Turkey.

Hitomi Tohyama, Shigeki Matsubara, Nobuo Kawaguchi, and Yasuyoshi Inagaki. 2005. Construction and utilization of bilingual speech corpus for simultaneous machine interpretation research. In *Proceedings of INTERSPEECH*, pages 1585–1588, Lisbon, Portugal.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of NIPS 2017*, pages 5998–6008, Long Beach, CA, USA.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-Sequence Models Can Directly Translate Foreign Speech. In *Proceedings of Interspeech 2017*, pages 2625–2629, Stockholm, Sweden.

Michael Melese Woldeyohannis, Laurent Besacier, and Million Meshesha. 2017. A Corpus for Amharic-English Speech Translation: the Case of Tourism Domain. In *Proceedings of ICT4DA 2017*, pages 129–139, Bahir Dar, Ethiopia.