MACHINE TRANSLATION VERBOSITY CONTROL FOR AUTOMATIC DUBBING

Surafel M. Lakew Marcello Federico Yue Wang† Cuong Hoang
Yogesh Virkar Roberto Barra-Chicote Robert Enyedi

Amazon AI

{surafelm|marcfede|yyuew|hoacuong|yvvirkar|rchicote|renyedi}@amazon.com

ABSTRACT

Automatic dubbing aims at seamlessly replacing the speech in a video document with synthetic speech in a different language. The task implies many challenges, one of which is generating translations that not only convey the original content, but also match the duration of the corresponding utterances. In this paper, we focus on the problem of controlling the verbosity of machine translation output, so that subsequent steps of our automatic dubbing pipeline can generate dubs of better quality. We propose new methods to control the verbosity of MT output and compare them against the state of the art with both intrinsic and extrinsic evaluations. For our experiments we use a public data set to dub English speeches into French, Italian, German and Spanish. Finally, we report extensive subjective tests that measure the impact of MT verbosity control on the final quality of dubbed video clips.

Index Terms— Machine Translation, Automatic Dubbing.

1. INTRODUCTION

Automatic Dubbing (AD) is the task of automatically replacing the speech in a video document with speech in a different language, while preserving as much as possible the user experience of the original video. AD differs from speech translation [1, 2] in significant ways. In speech translation, a speech utterance in the source language is recognized, translated (and possibly synthesized) in the target language. In speech translation, close to real-time response is expected and typical use cases include human-to-human interaction, traveling, live lectures, etc. Corresponding human tasks, from which data can be catered, are consecutive and simultaneous interpretation, where either isolated sentences or a continuous stream of speech are translated. On the other hand, AD tries to automate the localization of audiovisual content, a complex and demanding workflow [3] managed during post-production by dubbing studios.

A major requirement of dubbing is speech synchronization which, in order of priority, should happen at the utterance level (isochrony), lip movement level (lip synchrony) and body movement level (kinesic synchrony) [3]. Most of the work on AD [4, 5, 6], including this one, addresses isochrony, aiming to generate translations and utterances that match the phrase-pause arrangement of the original audio. Hence, given the transcript of a source speech utterance, provided with time stamps, the first step is to generate a translation that fits the duration of the original utterance [7, 8]. Then, a *prosodic alignment* [4, 5, 6] step follows, segmenting the translation into phrases and pauses corresponding to the phrases and pauses present in the original speech. Finally, the sequence of phrases and pauses is passed to a text-to-speech synthesizer that generates each phrase by adjusting the speaking rate to fit its required duration.

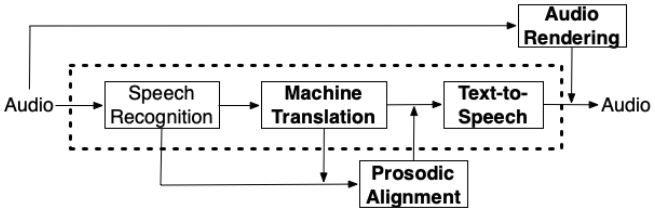

Fig. 1. Speech translation pipeline (dotted box) with enhancements introduced to perform automatic dubbing (in bold).

This paper focuses on the MT step, controlling the output length so that AD can produce utterances of the same duration of the original speech. In this work, we use the number of characters in the sentence as a proxy of the duration of its spoken realization. [1] Hence, we focus on controlling the number of characters in the MT output to be more or less the same as that of the source sentence. Our work is inspired by recent works that have addressed verbosity control for text summarization [9, 10, 11], of neural MT models [12, 13, 14, 7, 8] and ways to condition MT to side information [15, 16]. In particular, we implement and extend approaches proposed in [7, 8, 15, 16] to control the verbosity of MT for speech dubbing. As a significant difference to previous work we perform both intrinsic and extrinsic evaluations of the generated translations, for which we use a publicly available data set [17] with speeches from English to French, Italian, German and Spanish. Intrinsic evaluations measure MT quality and verbosity with respect to human post-edited translations matching length requirements, while extrinsic evaluations measure subjective quality of video clips dubbed by using the generated translations. To our knowledge, our work is the first to provide a systematic comparison of verbosity control methods in MT and a subjective evaluation that measures their usefulness on automatically dubbed content.

Our paper is arranged as follows. First, we describe the AD architecture used for our experiments. We then focus on existing and new methods for controlling the verbosity of MT output. Finally, we present and discuss experimental results of all compared methods.

2. DUBBING ARCHITECTURE

This work builds on the AD architecture presented in [5, 6] (Figure 1) that extends a speech-to-speech translation [1, 18, 19] pipeline with: neural machine translation (MT) robust to ASR errors and

†Author carried out the work during an internship at Amazon.

[1] We empirically found that characters work better for this purpose than syllables computed with the tool used in [7].

able to control verbosity of the output [20, 8, 21]; prosodic alignment (PA) [4] which addresses phrase-level synchronization of the MT output by leveraging the force-aligned source transcript; neural text-to-speech (TTS) [22, 23, 24] with precise duration control; and, finally, audio rendering that enriches TTS output with the original background noise (extracted via audio source separation with deep U-Nets [25, 26]) and reverberation, estimated from the original audio [27, 28].

3. MT WITH VERBOSITY CONTROL

Automatic dubbing calls for translations which can be fluently uttered within the same time interval of the original source speech [7, 8]. Given that text-to-speech can stretch its speaking rate without noticeable effects[2], ideally we would like MT to produce translations that are within a $\pm 10\%$ range of the original length, which we measure in number of characters. In the following, we present a range of approaches that we investigated to pursue this goal.

3.1. Naive Length Control

The simplest way to control verbosity of MT is to end the inference once the output has reached the target length. However, with the natural difference in verbosity among languages, it is obvious that this sole criterion could lead to poor MT performance. A better alternative is to leverage the already existing *length penalty* [29] used at search time to avoid the NMT model producing too short or incomplete translations. It normalizes the log-prob scoring function as:

$$S(t,s) = \frac{logP(t|s)}{LP(t)} + CP(s,t),$$

where the coverage penalty (CP) penalizes translation that fully covers the source, whereas length penalty (LP) is [29]:

$$LP(t) = (5+|t|)^{\alpha}/(5+1)^{\alpha}. \quad (1)$$

Following [8], we found that $\alpha = 0.5$ provides the best trade-off between verbosity and translation quality.

3.2. Verbosity Token

As proposed by [8], we can introduce a special source token that specifies the desired verbosity in the translation. To train NMT to learn this behavior, we first need to compute the target-source length ratio (LR) of all entries in the training data. Then, we categorize the training examples into three classes (*Short*, *Normal* and *Long*) based on their LR as follows:

$$v = \begin{cases} \texttt{Short} & \text{if } LR < 0.97 \\ \texttt{Normal} & \text{if } 0.97 \leq LR \leq 1.05 \\ \texttt{Long} & \text{if } LR > 1.05. \end{cases} \quad (2)$$

At training time, the verbosity token (v) is assigned to an embedding vector like any other token of the source vocabulary. Formally, we feed the MT encoder a sequence of embeddings as follows:

$$E_{source} = [E(v), E(tok_1), ..., E(tok_N)]. \quad (3)$$

Where $E(\dot{})$ is the embedding lookup function and N is the number of tokens in the source sentence. The model is trained end-to-end so that both $E(v)$ and all MT parameters are jointly learned.

[2]This holds at normal speaking rates, but clearly not when the original speech has already extreme speaking rates.

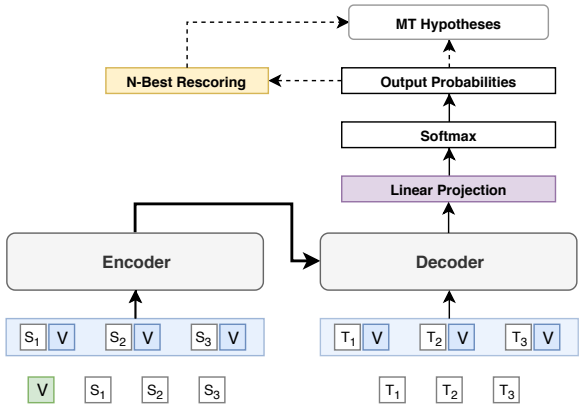

Fig. 2. Our MT modeling variants with verbosity control: (green) verbosity token (v) added to input source, (blue) summing v on the source/target embeddings, (violet) adding v as an additional bias and (yellow) rescoring output hypotheses.

At inference time, we prepend a desired v value to the source sequence (e.g. `Normal`). This encourages the MT model to generate translations that are within the corresponding LR range of Eq. 2.

Although the verbosity token approach has shown to be effective [8], we explore further means to inject verbosity information in the model and in the search process (see Figure 2).

3.3. Verbosity Embeddings

Verbosity information, once mapped to an embedding, can be integrated into the encoder and decoder in various ways.

Summing Verbosity to Token Embeddings

As an alternative to (3), we can feed the encoder with the sequence:

$$E_{source} = [E(tok_1) + E(v), ..., E(tok_N) + E(v)]. \quad (4)$$

Trivially, the same idea can be also applied to the input of the decoder. We will thus experiment with all three combinations: only encoder, only decoder, both encoder-decoder. Another alternative we consider, is to use the verbosity encoding and in addition the verbosity embedding in the decoder. Our motivation is to reinforce the influence of the verbosity token in MT and investigate whether this scheme makes any additional impact on controlling MT output verbosity without sacrificing translation quality.

Verbosity as Output Layer Bias

We can use the verbosity embedding $E(v)$ as an extra bias vector [16] in the final linear projection layer of the decoder:

$$O_t = WS_t + b + E(v), \quad (5)$$

where S_t is the decoder state at time t, W and b are the transformation and bias vector of the output layer and O_t is the output vector.

3.4. Fine-Tuning with Verbosity Information

It is suggested by [8] to apply training of the verbosity token as a fine-tuning stage of a pre-trained model. This work explores this direction further and compares two fine-tuning approaches:

- Single-stage fine-tuning: Fine-tune a generic model trained with large scale generic data with in-domain data augmented with verbosity token, as in [8].
- Two-stage fine-tuning: Fine-tune the generic model with verbosity information on generic data and then fine-tune again on in-domain data with verbosity information.

3.5. Rescoring Translation Output

In order to generate translations suited for dubbing, [7] proposed to rescore N-best translations (t) generated with a large beam size (B) with the following function:

$$S_d(t,s) = (1-\alpha)\log P(t \mid s) + \alpha S_p(t,s), \quad (6)$$

where S_p is the synchrony score computed by:[3]

$$S_p(t,s) = (1 + |len(t) - len(s)|)^{-1}. \quad (7)$$

The factor α is adjusted to set the relative importance of length-similarity versus translation-probability. As reported in [7] and confirmed by our experiments, for high α values the synchrony sub-score (S_p) can cause significant performance drop.

In this work, we hypothesize that it is suboptimal to use the synchrony sub-score as in Eq. 7 because it aims to make long output shorter and short output longer at the same time. This is not necessary because in practice, we often find the need of either reducing or increasing the length ratio (LR). More specifically, translation directions in our experiments are from *English* to other languages. To make MT output more or less the same as that of the source sentence, we often find the need of reducing the LR and not the other way around. To this end, we propose a new unidirectional version of the synchrony sub-score that encourages the decrease of the LR of the translations during rescoring:

$$S_p(t,s) = (1 + \frac{len(t)}{len(s)})^{-1}. \quad (8)$$

Compared to the original synchrony sub-score, our proposed scoring function is not only simpler but also fits better to the need of reducing the LR. As a side note, for the opposite translation directions (from other languages to *English*) we can simply reverse the LR of the translations during rescoring.

4. EXPERIMENTS

4.1. Data and Metrics

We evaluate the proposed approaches on four translation directions, English (En) to Italian (It), French (Fr), German (De) and Spanish (Es). We present results on the MuST-C corpus [17] of TED talks. To simulate realistic production settings, we also leverage proprietary parallel data at the magnitude of 10^7 sentences for MT model pre-training of each pair. Data is first preprocessed with scripts from the Moses [30] tool.[4] We then use SentencePiece [31][5] to learn a sub-word model with 32k merge operations, followed by segmentation of the detokenized text version.

To evaluate MT performance, we re-translated a subset of 620 sentences from the original MUST-C test, by asking our external vendors to produce translations close in length to the source. To measure MT quality, we use the BLEU score computed with SacreBLEU [32].[6] To evaluate verbosity control, we count the % of outputs meeting the target-source length ratio of $1 \pm 10\%$, which we consider acceptable for AD.

4.2. Model and Settings

All models are implemented using Transformer [20], with 6 layers of encoder-decoder network. Each layer constitutes sub-layers of self-attention of size 1024 with 16 heads and feed-forward with 4096 hidden dimensions. Adam optimizer [33] is used with a learning rate of 1×10^{-7}, that linearly increases for the first 4000 steps, followed by a decrease with an inverse square root of the model training steps. A dropout of 0.1 is applied on the attention layer, while 0.3 is used for the rest. All presented systems are trained on the large data set and fine tuned on the in-domain data. At fine-tuning time, employing similar settings has shown better performance, with the only exception of resetting the optimizer state. Training is performed on 8 V100 GPUs. Fine-tuning is performed for a total of 20 epochs. Models for evaluation are selected based on the validation loss. Except for rescoring models, where we set a beam size of 50 as in [7], we use a beam size of 5 for all inferences.

4.3. Intrinsic Evaluation

We first apply the single-stage fine-tuning method with our verbosity models. Experimental results with MT quality and verbosity control scores are in Figure 3. For all language pairs, the standard Transformer model (*Standard*) is presented as the reference system to be improved along both dimensions. Our observations are as follows.

First, the length penalty (*LenPenalty*) baseline model, representing a naive way of generating shorter translations, surprisingly shows a slight gain both in BLEU and verbosity control with respect to the *Standard*. Second, N-best rescoring as in [7] (*RescoreDiff*) displays mixed behaviours. Verbosity control on IT and ES reaches 90%, but BLEU drops, while on DE and FR it improves both dimensions. N-best rescoring using our proposed synchrony sub-score *RescoreRatio*, by contrast, shows a more consistent behaviour across all languages. For instance, it improves BLEU score and source-target length ratio for all language pairs with the `Normal` token.

Third, our verbosity models with `Short` and `Normal` tokens (*EncTok*) improve both metrics across all languages, except for $En-Es$ pair. Moreover, the `Short` token (marked) tends to have better performance than the `Normal` token (unmarked). This confirms the need of reducing the LR to make translation outputs from *English* to other languages more or less the same as that of the source sentence.

Moreover, in both `Short` and `Normal` settings, the verbosity embedding approaches (*DecEmb, EncDecEmb* and *EncTokDecEmb* show a less consistent behaviour. In particular, for DE and ES we observe MT quality drops by all such models, while for FR we observe MT quality drops by *EncDecEmb, EncTokDecEmb*. Finally, using the verbosity embedding as output layer bias in the decoder [16] did not provide any improvements in MT quality nor verbosity control with respect to the *Standard* model. For the sake of brevity they are not reported in the paper.

Given its consistent behavior we explore the *EncTok* method further. We first investigate the combination of the method and the two-stage fine-tuning procedure, as in Figure 3. By applying the two stage fine-tuning instead of the one-stage fine-tuning method, our verbosity token (*EncTokFT*) model with the `Normal` token gains

[3] With the minor difference that in [7] the length is expressed in syllables while here in characters.

[4] Moses: https://github.com/moses-smt/mosesdecoder

[5] https://github.com/google/sentencepiece

[6] SacreBLEU: https://github.com/mjpost/sacrebleu

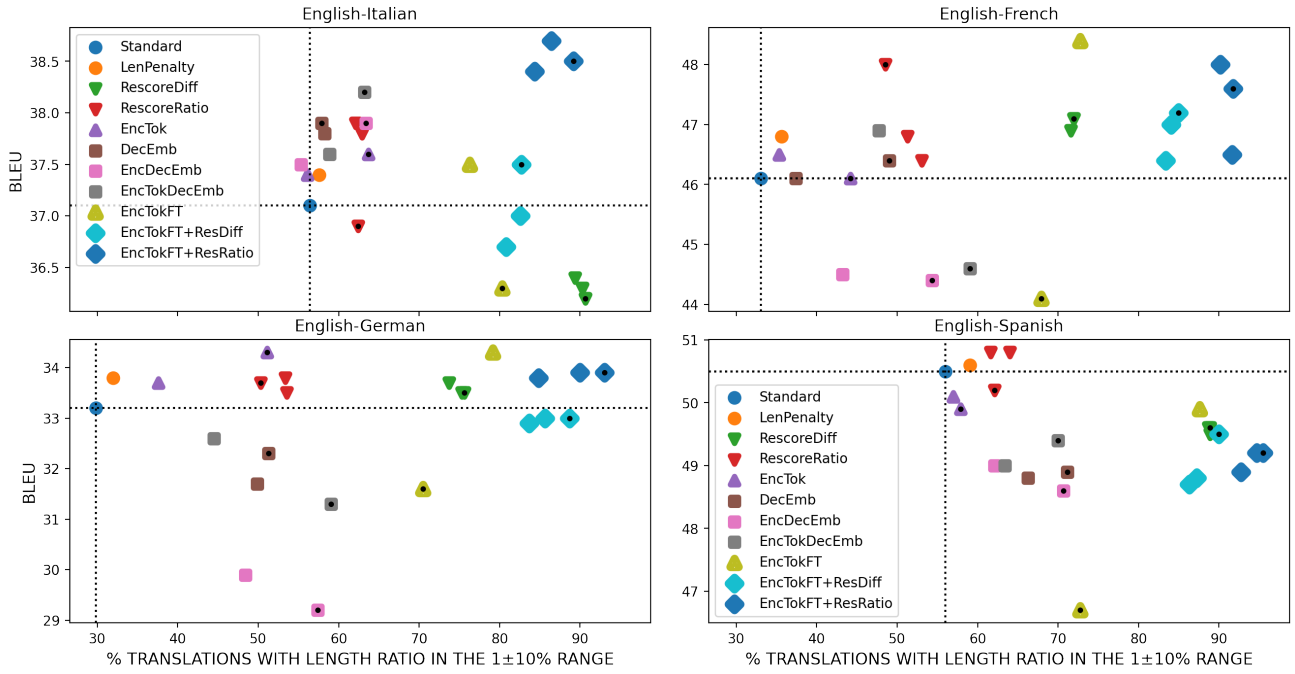

Fig. 3. Evaluation of verbosity control approaches in four translation directions. Methods are evaluated with respect to quality (BLEU) and verbosity (% of translations with acceptable length ratio). Dotted lines indicate performance of the standard MT model: hence, better systems should fall in the top right sector. Model families have distinct symbols: standard models (○), verbosity embedding (□), verbosity token (△), rescoring models (▽), verbosity token + rescoring (◇). Results of `Short` and `Normal` verbose/embedding models are shown, former are marked (•). Rescoring results are shown for limited sub-ranges of α, by marking (•) common α values for all languages.

better performance in all languages except Spanish where we see a slight loss in BLEU. In terms of verbosity control, the model translation outputs (unmarked) satisfy the length requirement well over 70% in all pairs. This verifies the effectiveness of our proposed fine-tuning procedure method in verbosity controlling. That is, the two-stage fine-tuning utilizes better the verbosity token in generating translation more or less the same as that of the source sentence.

Next, we combined the latter model with both N-best rescoring methods (*EncTokFT* + *RescoreDiff* and *EncTokFT* + *RescoreRatio*). We found the `Normal` setting works better for the combination and thus report only these results for the sake of clarity. Both combined methods further improve all previous methods. Moreover, rescoring using our poposed synchrony sub-score *RescoreRatio* is often better regarding both metrics of BLEU and source target length ratio. For instance *EncTokFT* + *RescoreRatio* pushes the percentage of acceptable length well over 90% for three pairs and to 89.9% for IT. In addition, *EncTokFT* + *RescoreRatio* improves BLEU score for all languages, but for Spanish (-2.5% relative BLEU): +4% for IT (+1.4 BLEU), +3% for FR (+1.5 BLEU) and +2% for DE (+0.70 BLEU).

4.4. Extrinsic Evaluation

We run an extrinsic subjective evaluation on a subset of 120 sentences to directly measure the impact that verbosity control of MT has on speech dubbing. In particular, we limit the comparison to Italian and German and to translations generated with *Standard* and our best model *EncTokFT* + *RescoreRatio*. After removing identical translations by the two systems, we end up with 100 and 110 sentences, respectively. We generate dubbed videos in Italian and German from them by using the architecture described in Section 2. As a reference, we also generate dubs from the reference translations. We split the test into batches that we assigned to a total of 40 subjects (proportions vary by language). Subjects were asked to watch the reference video and then rate their user experience with the two dubbing variants which were presented in random order and anonymously on a scale from 0 to 10 (higher is better).

We collected a total of 2,000 and 2,200 judgments for each variant and used them for a head-to-head comparison. By looking at the percentage of wins in Italian: *EncTokFT* + *RescoreRatio* got 38.7% wins against *Standard* got 32.45% ($p < 0.01$) (the rest were ties). For German, *EncTokFT* + *RescoreRatio* got 40.0% wins against *Standard* got 33.64% ($p < 0.02$) (the rest were ties).

5. CONCLUSIONS

We have presented and systematically compared verbosity control methods of MT in order to generate translations of length that is appropriate for automatic dubbing. Our analysis includes methods of integrating verbosity tokens and embeddings, fine-tuning strategies with verbosity information and finally rescoring functions to select outputs with the desired quality and verbosity. Compared to a standard Transformer MT model trained without verbosity information, our resulting best model not only produces translations much closer in length to the input, but often also better in translations. We also conducted a subjective evaluation on automatically dubbed videos using the translations generated by MT with and without verbosity control. The results confirm an increase in human preference for videos dubbed with the latter version.

6. REFERENCES

[1] F. Casacuberta, M. Federico, H. Ney, and E. Vidal, "Recent efforts in spoken language translation," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 80–88, 2008.

[2] M. Sperber and M. Paulik, "Speech Translation and the End-to-End Promise: Taking Stock of Where We Are," in *Proc. of the ACL*, pp. 7409–7421, July 2020.

[3] F. Chaume, "Synchronization in dubbing: A translation approach," in *Topics in Audiovisual Translation* (P. Orero, ed.), pp. 35–52, John Benjamins B.V., 2004.

[4] A. Öktem, M. Farrùs, and A. Bonafonte, "Prosodic Phrase Alignment for Machine Dubbing," in *Proc. Interspeech*, 2019.

[5] M. Federico, R. Enyedi, R. Barra-Chicote, R. Giri, U. Isik, A. Krishnaswamy, and H. Sawaf, "From Speech-to-Speech Translation to Automatic Dubbing," in *Proc. of IWSLT*, (Online), pp. 257–264, ACL, July 2020.

[6] M. Federico, Y. Virkar, R. Enyedi, and R. Barra-Chicote, "Evaluating and optimizing prosodic alignment for automatic dubbing," in *Proc. of Interspeech*, p. 5, 2020.

[7] A. Saboo and T. Baumann, "Integration of Dubbing Constraints into Machine Translation," in *Proc. of WMT*, (Florence, Italy), pp. 94–101, ACL, Aug. 2019.

[8] S. M. Lakew, M. Di Gangi, and M. Federico, "Controlling the output length of neural machine translation," in *Proc. IWSLT*, 2019.

[9] Y. Kikuchi, G. Neubig, R. Sasano, H. Takamura, and M. Okumura, "Controlling Output Length in Neural Encoder-Decoders," in *Proc. of EMNLP*, (Austin, Texas), pp. 1328–1338, ACL, Nov. 2016.

[10] A. Fan, D. Grangier, and M. Auli, "Controllable Abstractive Summarization," *Proc. of the 2nd Workshop on Neural Machine Translation and Generation*, Nov. 2017.

[11] S. Takase and N. Okazaki, "Positional Encoding to Control Output Sequence Length," *Proc. of NAACL*, Apr. 2019.

[12] J. Niehues, "Machine Translation with Unsupervised Length-Constraints," *Proc. of AMTA*, Apr. 2020.

[13] S. Agrawal and M. Carpuat, "Controlling Text Complexity in Neural Machine Translation," in *Proc. of EMNLP-IJCNLP*, (Hong Kong, China), pp. 1549–1564, ACL, Nov. 2019.

[14] K. Marchisio, J. Guo, C.-I. Lai, and P. Koehn, "Controlling the Reading Level of Machine Translation Output," in *Proc. of Machine Translation Summit XVII Volume 1: Research Track*, (Dublin, Ireland), pp. 193–203, Aug. 2019.

[15] C. D. V. Hoang, G. Haffari, and T. Cohn, "Improved Neural Machine Translation using Side Information," in *Proc. of the Australasian Language Technology Association Workshop 2018*, (Dunedin, New Zealand), pp. 6–16, Dec. 2018.

[16] P. Michel and G. Neubig, "Extreme adaptation for personalized neural machine translation," in *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, (Melbourne, Australia), pp. 312–318, Association for Computational Linguistics, July 2018.

[17] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "MuST-C: a Multilingual Speech Translation Corpus," in *Proc. NAACL*, pp. 2012–2017, 2019.

[18] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-Sequence Models Can Directly Translate Foreign Speech," in *Proc. Interspeech 2017*, pp. 2625–2629, ISCA, Aug. 2017.

[19] L. Cross Vila, C. Escolano, J. A. R. Fonollosa, and M. R. Costa-Jussà, "End-to-End Speech Translation with the Transformer," in *IberSPEECH 2018*, pp. 60–63, ISCA, Nov. 2018.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, pp. 5998–6008, 2017.

[21] M. Di Gangi, R. Enyedi, A. Brusadin, and M. Federico, "Robust neural machine translation for clean and noisy speech translation," in *Proc. IWSLT*, 2019.

[22] N. Prateek, M. Lajszczak, R. Barra-Chicote, T. Drugman, J. Lorenzo-Trueba, T. Merritt, S. Ronanki, and T. Wood, "In other news: A bi-style text-to-speech model for synthesizing newscaster voice with limited data," in *Proc. NAACL*, pp. 205–213, 2019.

[23] J. Latorre, J. Lachowicz, J. Lorenzo-Trueba, T. Merritt, T. Drugman, S. Ronanki, and K. Viacheslav, "Effect of data reduction on sequence-to-sequence neural TTS," in *Proc. ICASSP*, pp. 7075–7079, 2019.

[24] J. Lorenzo-Trueba, T. Drugman, J. Latorre, T. Merritt, B. Putrycz, R. Barra-Chicote, A. Moinet, and V. Aggarwal, "Towards Achieving Robust Universal Neural Vocoding," in *Proc. Interspeech*, pp. 181–185, 2019.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. ICMAI*, pp. 234–241, Springer, 2015.

[26] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," in *Proc. ISMIR*, 2017.

[27] H. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *Proc. IWAENC*, pp. 1–4, 2010.

[28] E. A. Habets, "Room impulse response generator," Tech. Rep. 2.4, Technische Universiteit Eindhoven, 2006.

[29] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[30] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, *et al.*, "Moses: Open source toolkit for statistical machine translation," in *Proc. of the ACL*, pp. 177–180, Association for Computational Linguistics, 2007.

[31] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," *arXiv preprint arXiv:1808.06226*, 2018.

[32] M. Post, "A call for clarity in reporting BLEU scores," in *Proc. of the Third Conference on Machine Translation: Research Papers*, (Belgium, Brussels), pp. 186–191, Association for Computational Linguistics, Oct. 2018.

[33] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.