Isochrony-Aware Neural Machine Translation for Automatic Dubbing

Derek Tam, Surafel M. Lakew, Yogesh Virkar, Prashant Mathur, Marcello Federico*

AWS AI Labs

{surafelm|yvvirkar|pramathu|marcfede}@amazon.com

Abstract

We introduce the task of isochrony-aware machine translation which aims at generating translations suitable for dubbing. Dubbing of a spoken sentence requires transferring the content as well as the speech-pause structure of the source into the target language to achieve audiovisual coherence. Practically, this implies correctly projecting pauses from the source to the target and ensuring that target speech segments have roughly the same duration of the corresponding source speech segments. In this work, we propose implicit and explicit modeling approaches to integrate isochrony information into neural machine translation. Experiments on English-German/French language pairs with automatic metrics show that the simplest of the considered approaches works best. Results are confirmed by human evaluations of translations and dubbed videos.

Index Terms: Machine Translation, Isochrony, Prosody, Verbosity, Automatic Dubbing

1. Introduction

Recent advancements in machine translation (MT), largely due to the success of transformer models, have improved the quality of MT significantly [1]. However, when MT is applied to specific use cases, like subtitle translation or automatic dubbing [2, 3, 4], translation quality is not the only dimension by which a model's performance is evaluated. In subtitles, translation of a source sentence should fit in a fixed block size [5]. Automatic Dubbing requires *isochrony* i.e. when the character is on-screen the translation (of source utterance line) should match the timing of original speech and speech-pause temporal arrangement in the original audio [6]. This means, pauses in the source utterance should be projected into the target translation in relatively similar positions [7]. In this paper, our focus is on translation and projection of pause markers in the correct position to **enable isochrony** in dubbing.

Currently in an automatic dubbing pipeline [2, 3, 4], a source utterance is first *translated* and then a prosodic alignment (PA) model [6] segments the translated text into phrases and pauses following the phrase-pause arrangement of the source utterance [4, 6]. [1] In these two steps, two distinct models are deployed, one for translation and one for segmentation, which is clearly a sub-optimal solution. Our hypothesis is that better and more suitable translations could be generated by taking into account the phrase-pause structure to be targeted.

In this paper, we propose to combine the two steps into a single MT model that directly generates translations including pause markers. We, therefore, introduce a new task of Isochrony-Aware MT (IAMT) where MT system should jointly transfer both meaning and the phrase-pause structure from source to target language.

The task of IAMT is challenging in different ways: 1) MT needs to learn two distinct modeling problems: MT and PA; 2) while learning to project pauses, MT should not deteriorate translation quality; 3) MT should temporally map a source segment (text between two pauses) into a target segment of similar duration. In particular, the third challenge requires MT to control the verbosity of translation at the segment level rather than just at the sentence level.

As part of recent efforts to achieve a better synchrony between source and target speech, most of the works have been focused on controlling the length of translated text i.e. its verbosity. [8] introduced a prefix verbosity control token to control for length and later [9] extended the same by generating multiple length controlled hypotheses and rescoring them according to a synchrony score [10]. [8, 11] controlled the verbosity by utilizing positional encoding in the transformer architecture [1] while [12] constrained the beam search to generate similar (source) length translations. With regards to synchronizing translation with speech, [4] introduced a prosodic alignment model and later [6] improved over that by utilizing speaking rate information and cross-lingual semantic matches to project source pauses to the target translation while [13] leverages the attention weights in neural MT. None of the previous works have looked at the problem of translation while maintaining speech synchronization as a whole except in a related work where [14] jointly learns to translate and project line breaks in the context of subtitling.

Despite the progress in adapting MT to use cases such as automatic dubbing, and subtitling, incorporating isochrony information in MT has not yet been explored. The main contributions of this work are:

- Introduce the task of IAMT, investigate several approaches and report experimental results on a publicly available speech translation data set.
- Introduce a suite of automatic metrics to jointly evaluate phrase-pause alignment and verbosity of the translated phrases with respect to the source.
- Run subjective human evaluations on different MT system outputs and the final dubbed videos to measure the impact of the proposed approaches.

2. Isochrony-Aware Machine Translation

The task of IAMT involves translating sentences in source language containing pause markers correctly to the target language, which includes 1) *projection of the pause markers* and 2) *verbosity control of phrases* (see Fig. 1). To incorporate variable speaking styles, we refer to phrase as the text between two pauses, and not necessarily a group of words acting as a grammatical unit.Below we will discuss our proposed approaches of implicitly and explicitly formulating IAMT.

*This work was done during an internship at Amazon.

[1] In this paper, following [6] we define pause as 300ms of silence between two consecutive spoken words. We define a phrase (or interchangeably a segment) as the text between two pauses.

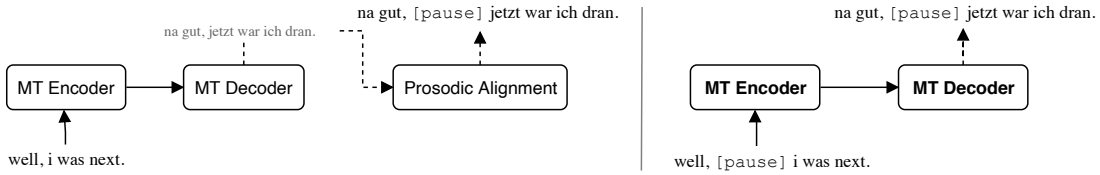

Figure 1: *Two step approach of translation and pause projection using MT and prosodic alignment [15] systems (left), and the proposed IAMT model generating translations along with* [pause] *markers denoting the temporal speech pause (right).*

Method	MT+PA	MT+[pause]	MT+DIS Emb	MT+DIS Att
SA	100	98.9	30.1	30.1

Table 1: *Preliminary results on En-De shows disentangled features (MT+DIS*) with embedding concatenation (Emb) or cross-attention (Att) under perform in terms SA metric.*

2.1. Implicit Control of Pause Marker positions

2.1.1. Pause Marker (MT+ [pause]*)*

Injecting meta information or linguistic markers in neural MT is a well studied topic [16, 17, 18]. A straightforward approach is to insert pause markers in the source and/or target text being generated. We simply add a [pause] token to delineate pauses in the source and target sentences. The MT model learns the phrase boundaries implicitly leveraging the [pause] positions along with tokens of the sequence. This way we incorporate pauses directly in the vocabulary of the model which allows it to learn semantics of the marker itself, and at the same time it implicitly learns to control the verbosity of the phrase by demarcating the phrase boundaries.

2.1.2. Disentangling Feature (MT+DIS)*

Factored MT has been used to inject external knowledge information in MT [19, 20] via a source factors [20, 21] and/or as a target factors [22]. This factorization allows the model to disentangle the meta information from the actual input. In our case, we add a binary feature in both source and target side as separate factors indicating whether there is a pause marker after the current token. The model then use these disentangled features to output two predictions: the next token, and whether there is a pause marker after the next token.

We experiment with two ways of modeling these two features: *i*) concatenating the embeddings of features and tokens in the input layer, and *ii*) having distinct encoders and decoders for tokens and binary features, connected by cross attention to model interactions between tokens and features.

We ran preliminary experiments with MT+[pause] and the disentangling approaches (see Table 1) against a strong baseline (MT+PA) where we first translate and then use a separate prosodic alignment model (with access to speech features) to project the pauses from source to target. These systems were evaluated on Over-/Under- Segmentation Accuracy (SA) i.e. accuracy of generating the same number of phrase segments in the target with respect to the source. SA is an initial indicator of the performance of the model on projection of pauses. To circumvent the lack of labeled training data with [pause] markers, we use a simulated training data for this process (refer Sec. 3.1 for more details). As it turns out, the segmentation accuracy goes down significantly for models with the disentangling feature, thus, we did not pursue this approach any further.

2.2. Explicit Control of Pause Marker positions

A drawback of the implicit approaches is the lack of control over the verbosity of each phrase between the pauses at inference time. While the projection of pause markers is important, we also need to control for the verbosity of phrases to achieve isochrony [23]. As a result, we further consider the objective of controlling the verbosity of each phrase and as a by product, we can also maximize the SA.

In previous work [24, 8], verbosity control is implemented using length-dependent positional encoding. Motivated by this, we look at the problem of transferring pause markers as a modeling problem. The main difference from their work is that in ours (*MT+LC*) we have to control for verbosity at phrase level.

Similar to [8], we first compute the ratio of number of characters left to generate in the target sequence: $1 - \frac{\text{\# char\_generated}}{\text{\# total\_char}}$ where $total\_char$ is the length of the target phrase. These floating point ratios on intervals of 0.1 from 0 to 1 are then quantized to an integer value between 0 and 10. The final embedding is the sum of token embeddings, sinusoidal positional embeddings [1], and the length dependent positional embeddings.

During training, the model uses the total number of characters in the reference target phrase to compute the ratio of characters left to generate. At inference time, where we do not have references, the model uses the total number of characters in the source phrase to compute the ratio of characters left to generate. Different from [24, 8] which stopped generating at end-of-sequence token, we stop generating when the ratio of characters left to generate is zero.

3. Experiments

3.1. Dataset

The HEROES data released by Oktem et. al. (2018) [13] is the only publicly available data for the task of IAMT for English-Spanish, however, it contains 7,000 samples in total which renders it rather small to train MT model. MuST-Cinema data [25] on the other hand has ≈200,000 samples for 7 language pairs but it is created specifically for speech to subtitles translation task, where segmentation is done on the basis of fixed length constraints (e.g. 42 characters), while for IAMT we require segmented input based on speech-pause information contained in the source audio.

MT model training with the proposed approaches require the source and target sentences to contain pause markers so that the model can learn to translate the content and project the pauses to the target side. However, since this information is unavailable from any of the open source data sets, we leverage the MuST-C speech-to-text translation data set [26], to create training and gold standard test sets.

3.2. Datasets for IAMT

3.2.1. Training Data

Training models for the IAMT task requires a much larger dataset (than the available ones) and performing forced alignment (for source pause markers) and post-editing (for target pause markers) at scale is very costly and time-consuming. To obtain source pause markers, a viable alternative is to use punctuation like comma or period characters. However, analyzing examples from *MuST-C-495* (refer Sec 3.2.2), we noticed that the speakers do not necessarily pause in correspondence of these punctuation characters or other linguistic cues in the text. Rather, they pause at any point of time - for instance to catch their breath or get a glass of water during their talk.

To generate the training data, we insert the pause markers in the source text in the following way: *First*, we computed a distribution of source phrase lengths from the source side of *MuST-C-495* set. *Second*, we randomly sampled a phrase length from this distribution and inserted pause markers after that desired phrase length. This way, we generated a phrase-pause structure in the source side of training data. To obtain target pause markers, we run a light weight PA module [6] (i.e. using only the cross-lingual semantic match features), to project pause markers from source to target text. In this way, for the *train/dev* data sets we collected about 200,000 sentence pairs with pause markers synthetically generated in both the source and target languages for two directions, English-German/French.

3.2.2. Evaluation Data

For evaluation, we collected unique pairs of 495 sentences from the official MuST-C test set (which contained duplicate sentence pairs). Given the corresponding audio, [6] annotated the pause information in the source sentence by force aligning the text with audio using Gentle aligner [27]. For each phrase in the source sentence, the corresponding target phrase was post-edited with human annotators to create a parallel data with phrase-pause structure where the target phrases were verbosity controlled i.e. similar in length to the source phrases.

3.3. Models

The baseline model (MT+PA) is a two step approach where we first translate the source text without pause marker using MT, and project the pause markers using the light weight PA module [6]. For IAMT task, we train models using the implicit (MT+`[pause]`) and the explicit (MT+LC) approaches proposed in Sec. 2.

Moreover, we compare the proposed models against the two step approach of Lakew et. al. [9]+PA approach, that trains MT with verbosity control, cascaded with the application of light-weight PA (as described in Sec 3.2.1). For all MT training we use the transformer base [1] model configuration.

3.4. Evaluation Metrics

Given that IAMT task is more complex than a standard translation task, we introduce additional metrics to measure three attributes: *i*) translation quality at phrase level, *ii*) segmentation accuracy, and *iii*) length compliance across source and target phrases.

We measure overall translation quality (at corpus level) using detokenized BLEU [28], while at phrase level we evaluate translation quality with ChrF score [29] (ChrF-Phrase) as precision for higher order n-grams might skew BLEU towards zero. To measure the accuracy of the projection of pauses over a data set, we compute the % of sentences for which the number of pauses in the target is the same as in the source (SA, segmentation accuracy). For verbosity control, to measure length compliance at the phrase level, we consider the % of sentences where length of every target phrase is within $\pm 10\%$ range in character count of the corresponding (order wise) source phrase (PhraseLC). Implicitly, PhraseLC also takes into account that number of pauses on either side should be the same.

Finally, we compute a single score, that gives an overall picture of the three attributes in question: $Acceptability = ChrF\text{-}Phrase * PhraseLC$.

	Method	BLEU	ChrF-Phrase	SA	PhraseLC	Acceptability
En-De	MT + PA	27.5	58.5	100	16.1	9.6
	MT + [pause]	27.8	59.5	99.8	19.7	11.7
	MT + LC	26.5	50.4	100	39.1	19.7
	Lakew et. al. [9]+PA	28.8	51.2	100	43	22
En-Fr	MT + PA	36.9	67.1	100	18.6	12.6
	MT + [pause]	38.0	68.8	96.3	20.1	13.8
	MT + LC	31.2	58.8	100	20.1	11.8
	Lakew et. al. [9]+PA	38.4	60	100	43.1	25.8

Table 2: *Results comparing the proposed IAMT approaches, MT+[pause] and* MT+LC *against the cascaded baseline MT+PA and the current best MT with verbosity control mechanism of Lakew et. al. [9]+PA, on MuST-C-495 test set [6].*

	Method	Acceptable	Fixable	Wrong
En-De	MT + [pause]	26.2	35.3	38.5
	MT + LC	12.1	29.0	58.9
	Lakew et. al. [9]	26.8	36.7	36.5
En-Fr	MT + [pause]	26.5	35.8	37.69
	MT + LC	8.2	28.2	63.6
	Lakew et. al. [9]	31.9	41.2	26.9

Table 3: *Human evaluation of MT system outputs (without pauses) on a* 200 *randomly selected unique samples from the post-edited benchmark.*

4. Results: Automatic Evaluation

Table 2 collects results for all systems evaluated on *MuST-C-495* post-edited test set for both En-De and En-Fr language pairs. Looking at the *Acceptability* scores, the approach of Lakew et. al. [9]+PA achieves the best results but this is expected as it first applies MT and a re-ranking module to control for verbosity, and then leverages a PA module specifically for phrase segmentation. Our aim is not to improve over this cascaded system, rather get as close as possible without deploying multiple modules into production.

The most interesting finding is that while MT+LC outperforms the MT+PA on length compliance of phrases it does so at trade-off with translation quality (c.f. ChrF-Phrase). This is expected because MT+LC optimizes on phrase level verbosity control. MT+`[pause]` on the other hand, consistently fares better against a strong baseline of MT+PA in terms of ChrF-Phrase, PhraseLC and Acceptability score. This means an implicit way of integrating pause markers into MT provides a better trade-off on all three attributes.

(I)		A	B	C	D	B′	D′
En-De	Smoothness	51.9	56.3	65.6	48.4	56.6	55.9
En-Fr	Smoothness	44.8	53.1	60.0	40.0	55.2	53.3

(II)		A vs. B		A vs. C		B vs. C		D vs. B		D′ vs. B′	
En-De	Wins	32.0	41.0*	48.4	30.8*	51.7	30.1*	34.8	40.9^{+}	37.4	37.5
En-Fr	Wins	36.9	38.2	61.4	25.8*	60.9	22*	29.9	43.4*	45.0	35.1*

Table 4: ***(I)*** *Automatic smoothness metric [6] and* ***(II)*** *Subjective user preferences (% of Wins) for automatic dubbing in a head to head comparison of: (A) MT+PA, (B) MT+[pause], (C) MT+*LC*, and (D) Lakew et al. [9]+PA. Models B′, D′ are versions of models B, D that also apply the relaxation mechanism in [6]. Significance testing is done for the Wins with levels $p < 0.05$ ($^{+}$) and $p < 0.01$ (*).*

5. Results: Human Evaluation

5.1. Machine Translation Evaluation

Human evaluation follows a simple yet an effective strategy to grade both quality and fluency of the MT outputs. Following [30], we ask subjects to rate 200 randomly selected translation subset of the test as acceptable, fixable, or wrong with respect to the reference.[2]

Table 3 show results comparing two of our proposed approaches against the state-of-the-art Lakew et al. [9] for MT verbosity control. For En-De, MT+`[pause]` shows comparable performance with acceptable translations at 26.2% with respect to [9] at 26.8%. For En-Fr, MT+`[pause]` drops by 5.4% from the best performing [9]. For the MT+LC model, we observed a large drop in the acceptable translations, which we regard as the outcome of an aggressive verbosity control that pushes the model to drop certain tokens. Overall, from the MT human evaluation we confirm that implicitly modeling the pause information with MT is a promising direction for Isochrony aware MT.

5.2. Dubbing Evaluation

We present results of human evaluation on a random subset of 50 single-sentence test videos. For each source sentence and corresponding video clip, we create dubbed videos using our dubbing architecture [4] with the following systems: (A) MT+PA, (B) MT+[`pause`], (C) MT+LC and (D) Lakew et al. [9] + PA. For the two systems in which no separate PA module is applied (B and C), time stamps of the source pauses are directly projected to the corresponding target pauses. For all systems, TTS audio is generated according to the duration of each target segment in order to fit the speech timing of the video. To reduce the cognitive load, we conduct separate evaluations comparing only two systems at a time. For all evaluations, we show as a reference a dubbed video generated from manually post-edited and segmented translations. Human subjects first watch the reference dubbed video and then rate viewing experience of videos dubbed with two systems on a scale of 0 to 10 with 10 being the highest quality and 0 being the worst quality.

We run evaluations on En-De and En-Fr directions with 40 human annotators, who are native speakers in the target language, with each of them grading 25 of the 50 videos, resulting in a total of 1,000 data points for each comparison. We report Wins, i.e., the % of times one system is preferred over the other.

Part **(I)** of Table 4 shows results for automatic evaluation with the Smoothness metric [6] that computes the stability of TTS speaking rate across contiguous target phrases. Part **(II)** shows the results for subjective human evaluation with the Wins metric. For both automatic and human metrics system B outperforms system A on both languages with relative improvements for Smoothness (De: +8.5%, Fr: +18.5%) and Wins (De: +28.1%, Fr: +3.5%) with statistically significant ($p < 0.01$) difference in Wins for De. B significantly outperforms C on both languages in terms of Wins (De: +57.1%, Fr: +137.9%). Though C has the better Smoothness compared to B, as shown in Sec. 5.1, C trades off on translation quality for improved Smoothness and hence results in automatically dubbed videos of lower quality.

5.2.1. Relaxation Mechanism

Comparing system D (with light weight PA, without speech features) with B, the latter is better on both Smoothness (De: +16.3%, Fr: +32.8%) and significant Wins (De: +17.5%, Fr: +45.2%) for both languages. However, this result does not take into account the relaxation mechanism [6] that can improve speaking rate smoothness.

Therefore, we applied the relaxation on the above two system and denote with B′ and D′ dubbing obtained with outputs from B and D after applying the relaxation. Note that, for D′ we apply a full fledged PA module (with speech features), while B′ is devoid of any such PA module. From Part **(I)**, we observe that Smoothness of D′ and B′ is improved as expected compared to D and B. Additionally Smoothness of D′ and B′ are now comparable. Also, D′ beats B′ for Wins on Fr (+28.2%, $p < 0.01$) and obtains comparable Wins for De. The reason is that after adding relaxation, while both systems reach comparable smoothness, D actually provides more acceptable translations than B (cf. Table 3). In fact, in order to generate high quality dubs, both translation quality and speaking rates are necessary components, and trading-off between these is the main challenge for IAMT.

6. Conclusion

In this work, we introduced an isochrony-aware MT task, where one has to transfer pause information from source to target along with translating the content. We proposed metrics to evaluate on multiple attributes; the correct number of pause markers, their positions, and verbosity at the level of phrase segments. We compared our proposed approaches (to model pause positions and translation) against strong baseline systems that decouples MT and prosodic alignment steps. We conducted automatic and human evaluations both on translation quality and on automatic dubbing, which relies on prosodic and temporal information projected from the source. As it turns out, the best approach to model both pause information and translation is to simply inject the pause markers in the text and let the model implicitly learn the two tasks.

[2] *Acceptable*: meaning is similar, fluency is good, *Fixable*: meaning is similar, fluency is poor, and *Wrong*: meaning is different.

7. References

[1] Ashish Vaswani, Noam Shazeer, Jakob Parmar, Nikia nd Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *31st Conference on Neural Information Processing Systems*, 2017.

[2] Alp Öktem, Mireia Farrùs, and Antonio Bonafonte, "Prosodic Phrase Alignment for Machine Dubbing," in *Proc. Interspeech*, 2019.

[3] Marcello Federico, Robert Enyedi, Roberto Barra-Chicote, Ritwik Giri, Umut Isik, Arvindh Krishnaswamy, and Hassan Sawaf, "From Speech-to-Speech Translation to Automatic Dubbing," in *Proc. of IWSLT*, Online, July 2020, pp. 257–264, ACL.

[4] Marcello Federico, Yogesh Virkar, Robert Enyedi, and Roberto Barra-Chicote, "Evaluating and optimizing prosodic alignment for automatic dubbing," in *Proc. of Interspeech*, 2020, p. 5.

[5] Alina Karakanta, Matteo Negri, and Marco Turchi, "Is 42 the Answer to Everything in Subtitling-oriented Speech Translation?," in *Proceedings of the 17th International Conference on Spoken Language Translation*, Online, July 2020, pp. 209–219, Association for Computational Linguistics.

[6] Yogesh Virkar, Marcello Federico, Robert Enyedi, and Roberto Barra-Chicote, "Improvements to prosodic alignment for automatic dubbing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[7] Marcello Federico, Yogesh Virkar, Robert Enyedi, and Roberto Barra-Chicote, "Evaluating and optimizing prosodic alignment for automatic dubbing," in *Interspeech*, 2020.

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

[9] Surafel M. Lakew, Marcello Federico, Yue Wang, Cuong Hoang, Yogesh Virkar, Roberto Barra-Chicote, and Robert Enyedi, "Machine translation verbosity control for automatic dubbing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[10] Ashutosh Saboo and Timo Baumann, "Integration of dubbing constraints into machine translation," in *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, Florence, Italy, Aug. 2019, pp. 94–101, Association for Computational Linguistics.

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

[12] Evgeny Matusov, Patrick Wilken, and Yota Georgakopoulou, "Customizing neural machine translation for subtitling," in *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, Florence, Italy, Aug. 2019, pp. 82–93, Association for Computational Linguistics.

[13] Alp Oktem, Mireia Farr, and Antonio Bonafonte, "Bilingual prosodic dataset compilation for spoken language translation," in *IberSPEECH*, 2018.

[14] Alina Karakanta, Matteo Negri, and Marco Turch, "Is 42 the answer to everything in subtitling-oriented speech translation?," in *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*, 2020.

[15] Marcello Federico, Robert Enyedi, Roberto Barra-Chicote, Ritwik Giri, Umut Isik, Arvindh Krishnaswamy, and Hassan Sawaf, "From speech-to-speech translation to automatic dubbing," in *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*, 2020.

[16] Rico Sennrich and Barry Haddow, "Linguistic input features improve neural machine translation," in *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, Berlin, Germany, Aug. 2016, pp. 83–91, Association for Computational Linguistics.

[17] Maria Nădejde, Siva Reddy, Rico Sennrich, Tomasz Dwojak, Marcin Junczys-Dowmunt, Philipp Koehn, and Alexandra Birch, "Predicting target language CCG supertags improves neural machine translation," in *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark, Sept. 2017, pp. 68–79, Association for Computational Linguistics.

[18] Cong Duy Vu Hoang, Gholamreza Haffari, and Trevor Cohn, "Improving neural translation models with linguistic factors," in *Proceedings of the Australasian Language Technology Association Workshop 2016*, Melbourne, Australia, Dec. 2016, pp. 7–14.

[19] Mercedes García-Martínez, Loïc Barrault, and Fethi Bougares, "Factored neural machine translation," *CoRR*, vol. abs/1609.04621, 2016.

[20] Patrick Wilken and Evgeny Matusov, "Novel applications of factored neural machine translation," *CoRR*, vol. abs/1910.03912, 2019.

[21] Georgiana Dinu, Prashant Mathur, Marcello Federico, Stanislas Lauly, and Yaser Al-Onaizan, "Joint translation and unit conversion for end-to-end localization," in *Proceedings of the 17th International Conference on Spoken Language Translation*, Online, July 2020, pp. 265–271, Association for Computational Linguistics.

[22] Aleš Tamchyna, Marion Weller-Di Marco, and Alexander Fraser, "Modeling target-side inflection in neural machine translation," in *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark, Sept. 2017, pp. 32–42, Association for Computational Linguistics.

[23] Mayank Sharma, Yogesh Virkar, Marcello Federico, Roberto Barra-Chicote, and Robert Enyedi, "Intra-Sentential Speaking Rate Control in Neural Text-To-Speech for Automatic Dubbing," in *Proc. Interspeech 2021*, 2021, pp. 3151–3155.

[24] Sho Takase and Naoaki Okazaki, "Positional Encoding to Control Output Sequence Length," *Proc. of NAACL*, Apr. 2019.

[25] Alina Karakanta, Matteo Negri1, and Marco Turchi1, "Must-cinema: a speech-to-subtitles corpus," in *The 12th Conference on Language Resources and Evaluation (LREC)*, 2020.

[26] Roldano Cattoni, Mattia Antonino, Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi, "Must-c: A multilingual corpus for end-to-end speech translation," in *Computer Speech and Language Journal*, 2020.

[27] R. M. Ochshorn and M. Hawkins, "Gentle Forced Aligner," 2017.

[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[29] Maja Popović, "chrF: character n-gram F-score for automatic MT evaluation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, Sept. 2015, pp. 392–395, Association for Computational Linguistics.

[30] Surafel M. Lakew, Yogesh Virkar, Prashant Mathur, and Marcello Federico, "Isometric mt: Neural machine translation for automatic dubbing," *arXiv preprint arXiv:2112.08682*, 2021.