IMPROVEMENTS TO PROSODIC ALIGNMENT FOR AUTOMATIC DUBBING

Yogesh Virkar Marcello Federico Robert Enyedi Roberto Barra-Chicote

Amazon

ABSTRACT

Automatic dubbing is an extension of speech-to-speech translation such that the resulting target speech is carefully aligned in terms of duration, lip movements, timbre, emotion, prosody, etc. of the speaker in order to achieve audiovisual coherence. Dubbing quality strongly depends on isochrony, i.e., arranging the translation of the original speech to optimally match its sequence of phrases and pauses. To this end, we present improvements to the prosodic alignment component of our recently introduced dubbing architecture. We present empirical results for four dubbing directions – English to French, Italian, German and Spanish – on a publicly available collection of TED Talks. Compared to previous work, our enhanced prosodic alignment model significantly improves prosodic alignment accuracy and provides segmentation perceptibly better or on par with manually annotated reference segmentation.

Index Terms— speech translation, text to speech, 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 possibly the user experience of the original video. AD dubbing differs from speech translation [1, 2, 3, 4] 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. On the other hand, AD tries to automate the localization of audiovisual content, a complex and demanding work flow [5] 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 (kinetic synchrony) [5]. Most of the work on AD [6, 7, 8], including this one, addresses isochrony, which aims to generate translations and utterances that match the phrase-pause arrangement of the original audio. Given a source sentence transcript, the first step is to generate a translation of more or less the same "duration" [9, 10], e.g. number of characters or syllables. The second step, called *prosodic alignment* (PA) [6, 7, 8], segments the translation into phrases and pauses of the same duration of the original phrases.

This paper focuses on the PA step, by comparing previous [6, 7, 8] and new methods that allow to optimally segment and temporally align a translation with the original phrases. Differently from previous work, we perform intrinsic and extrinsic evaluations of PA an a significantly larger test set extracted from the MUST-C corpus [11] and on four dubbing directions, English (en) to French (fr), Italian (it), German (de) and Spanish (es). Intrinsic evaluations measure the accuracy, fluency and smoothness of PA with respect to manually post-edited and segmented translations, while extrinsic evaluations measure subjective quality of video clips dubbed by applying different PA models on the human translations.

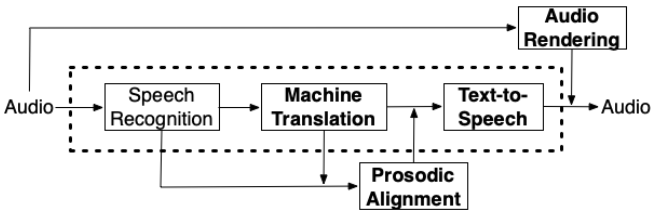

Fig. 1. Speech translation pipeline (dotted box) with enhancements introduced to perform automatic dubbing (in bold).

Our paper is arranged as follows. First, we describe the automatic dubbing architecture used for our experiments, then, we focus on existing and new PA methods, and finally we present and discuss experimental results of all compared methods.

2. DUBBING ARCHITECTURE

We build on the automatic dubbing architecture presented in [7, 8], and described in Figure 1, that extends a speech-to-speech translation [1, 2, 3] pipeline with: neural machine translation (MT) robust to ASR errors and able to control verbosity of the output [12, 10, 13]; prosodic alignment (PA) [6] which addresses phrase-level synchronization of the MT output by leveraging the force-aligned source transcript; neural text-to-speech (TTS) [14, 15, 16] 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 [17, 18]) and reverberation, estimated from the original audio [19, 20].

3. RELATED WORK

In the past, there has been little work to address prosodic alignment for automatic dubbing [6, 7, 8]. The work of [6] utilized the attention mechanism of neural machine translation to achieve isochrony. While this approach achieves linguistic similarity between corresponding source and target phrases, it has no mechanism to explicitly control for uneven or extreme speaking rates that can cause unnatural sounding dubbing. This was partly addressed in [7] by segmenting the translation according to the length similarity between corresponding source-target phrases. Moreover, [7] proposed a more efficient implementation based on dynamic programming as opposed to generating and rescoring segmentation hypotheses [6]. More recently, [8] further improved on speech fluency by not only controlling the speaking rate match between corresponding source-

To appear in *Proc. ICASSP2021, June 06-11, 2021, Toronto, Canada*

target phrases but also the speaking rate variation across consecutive target phrases. Additionally, it introduced a mechanism to relax the timing constraints to cope with too high TTS speaking rates. While it achieves much smoother speech, the human evaluation of the automatically dubbed videos revealed that users still prefer human-annotated reference segmentation over the one produced by the model. Finally, another limitation of this study is the use of a rather small test set. In this work, we address these issues by using a much larger dataset and combine the advantages of content-based [6] and fluency-based [7, 8] approaches. In sec. 6, we provide a direct comparison of our present work with [8] and an indirect comparison of our work with [6].

4. PROSODIC ALIGNMENT

The goal of the PA [6, 7, 8] is to segment the target sentence to optimally match the sequence of phrases and pauses in the source utterance. Let $\mathbf{e} = e_1, e_2, \ldots, e_n$ be a source sentence of n words, segmented according to k breakpoints $1 \leq i_1 < i_2 < \ldots i_k = n$, denoted with $\mathbf{i}$. Let the temporal duration of $\mathbf{e}$ be T and the temporal intervals of the segmentation $\mathbf{i}$ be $s_1 = [l_1, r_1], \ldots, s_k = [l_k, r_k]$, denoted by $\mathbf{s}$, s.t. $l_1 \geq \Delta\epsilon$, $l_i < r_i$, $l_{i+1} - r_i \geq \Delta\epsilon$, $T - r_k \geq \Delta\epsilon$, where $\Delta\epsilon$ is the minimum silence interval after (and before) each break point.[1] Given a target sentence $\mathbf{f} = f_1, f_2, \ldots, f_m$ of m words, the goal is to find k breakpoints $1 \leq j_1 < j_2 < \ldots j_k = m$ (denoted with $\mathbf{j}$) that maximize the probability:

$$\max_{\mathbf{j}} \log \Pr(\mathbf{j} \mid \mathbf{i}, \mathbf{e}, \mathbf{f}, \mathbf{s}) \tag{1}$$

By assuming a Markovian dependency on $\mathbf{j}$, i.e.:

$$\Pr(\mathbf{j} \mid \mathbf{i}, \mathbf{e}, \mathbf{f}, \mathbf{s}) = \sum_{t=1}^{k} \log \Pr(j_t \mid j_{t-1}; t, \mathbf{i}, \mathbf{e}, \mathbf{f}, \mathbf{s}) \tag{2}$$

and omitting from the notation the constant terms $\mathbf{i}$,$\mathbf{e}$,$\mathbf{f}$ and $\mathbf{s}$ we derive the following recurrent quantity:

$$Q(j, t) = \max_{j' < j} \log \Pr(j \mid j'; t) + Q(j', t-1) \tag{3}$$

where $Q(j, t)$ denotes the log-probability of the optimal segmentation of $\mathbf{f}$ up to position j with t break points. The implicit assumption is that corresponding source and target segments, defined by $\mathbf{i}$ and $\mathbf{j}$, have exactly the same duration (isochrony), defined by $\mathbf{s}$. In [8], we allow target segments to possibly extend the source interval by some fraction of $\Delta\epsilon$ to the left and to the right, which we call δ_l and δ_r. In this work, we additionally allow trimming by having negative relaxations such that $\delta_l, \delta_r \in \left\{0, \pm\frac{1}{4}, \pm\frac{2}{4}, \pm\frac{3}{4}, \pm 1\right\}$. Hence, we tradeoff isochrony for the flexibility of adjusting speaking rates to improve the viewing experience. Thus, we optimize

$$Q\left(j, \delta_l, \delta_r; t\right) = \max_{j' < j \,:\, \delta_r' \leq 1 - \delta_l} \log \Pr\left(j, \delta_l, \delta_r \mid j', \delta_l', \delta_r'; t\right) + Q\left(j', \delta_l', \delta_r'; t-1\right) \tag{4}$$

Here Q is the score of the optimal segmentation into t segments up to position j, with relaxations δ_l, δ_r on the last segment. Hence, not only different breakpoints j for the t-segment are evaluated, but also relaxations of the original time interval $s_t = [l_t, r_t]$, to the right by $\delta_r \Delta\epsilon$ and to the left by $\delta_l \Delta\epsilon$. We denote the relaxed interval by s_t^*. The constraint $\delta_r' \leq 1 - \delta_l$ in (4) makes sure that the left relaxation of segment t does not overlap with the right relaxation of segment $t-1$. Additionally, owing to negative relaxations, we note that we have implicit though trivial restrictions on the choice of $\delta_l, \delta_r, \delta_r', \delta_l'$ to ensure $s_t^* > 0, s_{t-1}^* > 0$.

We define the model probability with a log-linear model:

$$\log \Pr\left(j, \delta_l, \delta_r \mid \ldots; t\right) \propto \sum_{k=1}^{5} w_a \log s_a\left(j, \delta_l, \delta_r, \ldots; t\right) \tag{5}$$

where weights w_a are learned from data and functions s_a model the following features:

1. Language model score of target break point.
2. Cross-lingual semantic match score across source and target segments.
3. Speaking rate variation across target segments.
4. Speaking rate match across source and target segments.
5. Isochrony score for left and right relaxations.

Speaking rate computations rely on the strings $\tilde{f}_t$ and $\tilde{e}_t$, denoting the t-th source and target segments, as well as the original interval s_t and the relaxed interval s_t^*. Hence, the speaking rate of a source (target) segment is computed by taking the ratio between the duration of the utterance by source (target) TTS run at normal speed and the source (target) interval length,[2] i.e:

$$r_e(t) = \frac{\texttt{duration}(\texttt{TTS}_e(\tilde{e}_t))}{|s_t|} \tag{6}$$

$$r_f(t) = \frac{\texttt{duration}(\texttt{TTS}_f(\tilde{f}_t))}{|s_t^*|} \tag{7}$$

4.1. Language model

As in previous work [7, 8], we define a language model score that estimates the probability of placing a break between consecutive target words f_{t-1} and f_t:

$$s_{lm}(j, f_{t-1}, f_t) = \Pr\left(br | f_{t-1}, f_t\right) = \frac{c\left(g_{t-1}, br, g_t\right)}{c\left(g_{t-1}, br, g_t\right) + c\left(g_{t-1}, g_t\right)} \tag{8}$$

To use this feature, we first map the sentence $\mathbf{f}$ into a sequence $\mathbf{g}$ of parts-of-speech[3], where punctuation marks denoting pauses (period,comma, colon or semicolon) are mapped to the *br* class. Unlike our previous work, we now compute the probability directly from counts extracted from the training portion of the MUST-C corpus.

4.2. Cross-lingual semantic match

The pioneering work of [6] exploits the attention mechanism in neural machine translation to segment and align the translation to the source phrases. To capture such semantic similarity between corresponding source and target phrases, we define the cross-lingual semantic match $s_{cm}(\cdot)$ as

$$s_{cm}\left(\tilde{e}_t, \tilde{f}_t; t\right) = \cos\left(\phi\left(\tilde{e}_t\right), \phi(\tilde{f}_t)\right) \tag{9}$$

where $\phi(\cdot)$ denotes the encoding of the input phrase by using a pre-trained multilingual sentence embedding model [23]. In Sec. 6.2,

[1] In this work the minimum silence interval $\Delta\epsilon$ is set to $300ms$.

[2] We run TTS on the entire sentence, force-align audio with text [21, 22] and compute segment duration from the time-stamps of the words.

[3] We use https://aws.amazon.com/comprehend for this step.

we provide a quantitative comparison of using different multilingual embedding models. Note that the accuracy of the prosodic alignment depends strongly on matching the linguistic content between the source and target phrases and our past work [8] shows that accuracy impacts strongly the subjective viewing experience. Hence, we expect to benefit by using such content-matching feature.

4.3. Speaking rate variation

As in [8] we penalize hypotheses resulting in high variation in speaking rates for consecutive target phrases, we define the speaking rate variation score $s_{sv}(\cdot)$ as follows:

$$s_{sv}\left(\tilde{f}_t, s_t^*, \tilde{f}_{t-1}, s_{t-1}^*; t\right) = 1 - \frac{|r_f(t) - r_f(t-1)|}{r_f(t-1)} \quad . \tag{10}$$

Naturally, this feature works for $t \geq 2$ and reaches its maximum value when consecutive phrases have the same speaking rate.

4.4. Speaking rate match

We realized that the speaking rate match score in [8] does not take into account global information of the target sentence, such as its verbosity relative to the source sentence. Hence, to better match speaking rates between target and source segments, we introduce the factor:

$$\beta = \frac{\texttt{duration}(\texttt{TTS}_f(\mathbf{f}))}{\texttt{duration}(\texttt{TTS}_e(\mathbf{e}))} \tag{11}$$

that computes the ratio of the duration of the target sequence $\mathbf{f}$ and the source sequence $\mathbf{e}$, both synthesized using TTS at normal speaking rate. Thus, we define the speaking rate match score as:

$$s_{sm}\left(\tilde{e}_t, \tilde{f}_t, s_t, s_t^*; t\right) = 1 - \frac{|r_f(t) - \beta r_e(t)|}{\beta r_e(t)} \quad . \tag{12}$$

This feature reaches it maximum when the target speaking rate is identical to the scaled source speaking rate.

4.5. Isochrony score

We extend the isochrony score $s_{is}(\cdot)$ of [8] to positive and negative relaxations δ_l, δ_r, as:

$$s_{is}(\delta_l, \delta_r) = 1 - [\alpha \ |\delta_l| + (1-\alpha) \ |\delta_r|] \tag{13}$$

This feature reaches its maximum when no relaxation occurs ($\delta_r = \delta_l = 0$), that is when the TTS output is stretched to exactly fit the duration of the original utterance. The relaxation mechanism of [8] is able to mitigate only very high speaking rates. To mitigate very low speaking rates, we introduce negative values for relaxation and hence use absolute values for δ_l, δ_r in (13). Since relaxations are less tolerated at the beginning than at the end of a phrase [5], we set $\alpha > \frac{4}{5}$ such that left relaxation is always more penalized than the right, i.e.:

$$\alpha\,|\delta_l| > (1-\alpha)\,|\delta_r| \quad \forall \delta_l, \delta_r \in \left\{0, \pm\frac{1}{4}, \pm\frac{2}{4}, \pm\frac{3}{4}, \pm 1\right\} \tag{14}$$

5. EVALUATION DATA

To train and evaluate PA, we re-translated and annotated a total of 495 video clips from 20 TED talks of the MUST-C corpus such that a clip contains a single sentence with at least one pause of at least 300ms. Using [21], we time aligned the English text with the audio. Using external vendors, we manually adapted and segmented the available translations in 4 languages - French, German, Italian and Spanish - so as to fit the duration and segmentation of corresponding English utterances. For automatic evaluation, we use the metrics Accuracy, Fluency and Smoothnes as defined in [8].

6. EXPERIMENTS

To test the importance of relaxation mechanism, we evaluate our new PA model without (B) and with (C) relaxations and compare our results with the best model with relaxation from our previous work [8], i.e. model A. To simplify search and run ablation tests across models, we find optimal feature weights using hierarchical grid search with convex combinations of feature pairs [4]:

A: $\left(s_{is}, (s_{lm}, (s_{sm}, s_{sv})_{w_{sm}})_{w_{lm}}\right)_{w_{is}}$

B: $\left(s_{cm}, (s_{lm}, (s_{sm}, s_{sv})_{w_{sm}})_{w_{lm}}\right)_{w_{cm}}$

C: $\left(s_{is}, \left(s_{cm}, (s_{lm}, (s_{sm}, s_{sv})_{w_{sm}})_{w_{lm}}\right)_{w_{cm}}\right)_{w_{is}}$

Note that for model A, the features s_{is}, s_{lm} and s_{sm} are defined differently than for models B and C presented in this paper.

6.1. Two-step optimization

As we explain in Sec. 6.2, performance of model A [8] on the expanded test dataset of 495 sentences is significantly lower than that on the previous corpus. In particular, we found that relaxations do not help to improve accuracy but only improve fluency and smoothness. Hence, to find the optimal feature weights for model C we decided to utilize a two-step optimization procedure. In the first step, we find optimal weights w_{sm}, w_{lm}, w_{cm} by maximizing average accuracy. This is equivalent to training model B. In the second step, we find the optimal weight w_{is} by maximizing average smoothness by keeping the segmentation obtained in step 1. Thus, model C has the same segmentation (and hence accuracy) of model B but should have better smoothness.

6.2. Automatic evaluation

Table 1 shows the results of automatic evaluation. All the observed improvements of the new models B and C on our previous work A are statistically significant [24]. For all four languages, model B outperforms model A with relative improvements in accuracy – fr: +99.4%, it: +87.3%, de: +80.2%, es: +86.1%. This shows that the improvements in the speaking rate match feature, the language model and the addition of cross-lingual semantic match feature provides a strong improvement in the scoring function. The reduction in fluency and smoothness metrics can be attributed to the fact that A uses relaxations and B does not.

Table 3 shows a comparison of accuracy for PA model using only the cross-lingual semantic match feature with different pretrained models such as multilingual universal sentence encoder

[4] Where $(a, b)_\theta := \theta a + (1-\theta) b$ with $\theta \in [0.1]$

	Automatic	A	B	C	R
en-fr	Accuracy	35.6%	**70.9%***	–	100%
	Fluency	**87.9%**	62.8%*	82.6%*	59.2%
	Smoothness	**82.4%**	68.5%*	81.0%*	64%
en-it	Accuracy	43.0%	**80.6%***	–	100%
	Fluency	**78.4%**	50.7%*	59.6%*	52.3%
	Smoothness	**82.9%**	67.7%*	73.8%*	67.6%
en-de	Accuracy	39.8%	**71.7%***	–	100%
	Fluency	68.1%	57.6%*	**73.5%***	59.1%
	Smoothness	70.2%	65.0%*	**74.0%***	64.5%
en-es	Accuracy	43.4%	**80.8%***	–	100%
	Fluency	**77.0%**	44.9%*	52.1%*	44.2%
	Smoothness	**77.4%**	68.0%*	75.1%*	68.2%

Table 1. Results of automatic evaluation with prosodic alignments: (A) previous work [8], (B) new model without relaxation, (C) new model with relaxation and manual reference (R). Test set is made of 495 sentences. Significance testing is against model A, with levels $p < 0.05$ ($^+$) and $p < 0.01$ (*). Best PA results are in bold face.

	Manual	A vs. C		C vs. R	
en-fr	Wins	22.6%	52.4%*	39.6%	30.2%*
	Score	4.5	5.14*	4.96	4.76*
en-it	Wins	26.9%	36.8%*	28.3%	27.9%
	Score	4.55	4.76 *	4.49	4.49
en-de	Wins	27.3%	47.6%*	38.8%	31.8%*
	Score	4.81	5.38*	5.24	5.15
en-es	Wins	24.6%	34.5%*	20.8%	23.4%
	Score	4.56	4.78*	5.05	5.07

Table 2. Results of manual evaluations with prosodic alignments: (A) previous work [8], (C) new model with relaxation and manual reference (R). Test set is made of 50 video clips. Significance testing is with levels $p < 0.05$ ($^+$) and $p < 0.01$ (*).

(mUSE) [25], sentence BERT (SBERT) [26], language agnostic bert sentence embeddings (LaBSE) [27] and language agnostic sentence representations (LASER) [23]. LASER substantially outperforms all other models on all languages for our dataset and is hence used for s_{cm}. This configuration can be seen as a proxy of the PA model of [6], although we use a pretrained state-of-the-art embedding model rather than attention-based model and implement PA with dynamic programming.

The full model C improves on model B's Fluency (fr: +31.5%, it: +17.5%, de: +27.7%, es: +16.2%) and Smoothness (fr: +18.2%, it: +9%, de: +13.9%, es: +10.4%) without sacrificing on B's accuracy thanks to the two-step optimization procedure described in Sec. 6.1. Though not shown here, we note that compared to using positive relaxations, the addition of negative relaxations improves fluency and smoothness on average across all languages by +9% and +5% respectively. In comparing models C and A, we see a drop in Fluency in all languages except German (fr: -5.9%, it: -23.9%, de: +8%, es: -32.3%), as well as in Smoothness (fr: -1.7%, it: -11%, de: +5.5%, es: -2.9%). The observed drop in these metrics can be attributed to the restriction on the choice of segmentation that the two-step optimization imposes on C. In the next section, we show however that the overall dubbing quality of model C is far superior to that of model A, thanks to the higher accuracy which out-weighs the lower fluency and smoothness.

	mUSE	SBERT	LaBSE	LASER
en-fr	46.06%	51.11%	50.3%	**61.21**%
en-it	59.19%	63.84%	67.27%	**74.95**%
en-de	54.14%	57.17%	59.19%	**66.06**%
en-es	61.01%	64.04%	67.68%	**74.55**%

Table 3. Model accuracy using only the cross-lingual semantic match feature on pretrained models of mUSE [25], SBERT [26], LaBSE [27] and finally LASER [23] that significantly outperforms all other models on our dataset.

6.3. Human evaluation

For comparability with our previous work, we present results of human evaluation on the same subset of 50 test sentences used in [8]. For each language and each test sentence, starting with manually post-edited translation, we apply PA with models A and C, followed by neural text-to-speech and audio rendering steps as described in Sec. 2 to generate the dubbed videos. As reference (R), we also dubbed videos using the manual segmentation. We asked native speakers in each language to grade the viewing experience of each dubbed video on a scale of 0-10. To reduce the cognitive load, we perform two distinct evaluations, comparing two conditions in each case: A vs. C and C vs. R. We run evaluations using Amazon Mechanical Turk with 20 subjects each grading all videos for a total of 2000 scores in each language.

We compare A, C and R using the Wins (percentage of times one condition is preferred over the other) and Score (average score of dubbed videos) metrics. We use the *linear mixed-effects model*[5] (LMEM) by defining subjects and sentences as random effects [29].

Table 2 shows that C clearly outperforms A on all languages for both Wins (fr: +131.8%, it: +36.8%, de: +74.4%, es: +40.2%) and Score (fr: +14.2%, it: +4.6%, de: +11.9%, es: +4.8%) with all results statistically significant ($p < 0.01$). Comparing C with R, for it and es languages, we find no statistically significant difference for either metric. Instead, for de and fr we find +18% (+1.7%) and +23.7% (+4.2%) relative gains in Wins (Score), respectively. To understand why C outperforms R on fr and de, we used a LMEMs to explain observed Score variations by means of the three automatic metrics. We observe that for fr, smoothness is the only statistically significant factor ($p < 0.01$), while for de all three metrics are significant ($p<0.01$) in the following order of importance: smoothness, fluency, accuracy. This analysis confirms the importance of the relaxation mechanism and the two-step optimization process that help improve smoothness and fluency without sacrificing accuracy.

7. CONCLUSIONS

In this work, we presented the improved prosodic alignment component for automatic dubbing and evaluated it on a significantly larger test set with support for four dubbing directions. Modifications to the language model and speaking rate match features coupled with the addition of cross-lingual semantic match feature improves significantly the accuracy. Our two-step optimization process with the addition of negative relaxations helps improve smoothness and fluency without degrading accuracy. From the perspective of dubbing experience our new model provides segmentation vastly superior to our previous work and perceptibly better or on par compared to the reference segmentation for all four languages.

[5]We used the `lme4` package for R [28].

8. 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] Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen, "Sequence-to-Sequence Models Can Directly Translate Foreign Speech," in *Proc. Interspeech 2017*. Aug. 2017, pp. 2625–2629, ISCA.

[3] Laura Cross Vila, Carlos Escolano, Jos A. R. Fonollosa, and Marta R. Costa-Juss, "End-to-End Speech Translation with the Transformer," in *IberSPEECH 2018*. Nov. 2018, pp. 60–63, ISCA.

[4] Matthias Sperber and Matthias Paulik, "Speech Translation and the End-to-End Promise: Taking Stock of Where We Are," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020, pp. 7409–7421, Association for Computational Linguistics.

[5] Frederic Chaume, "Synchronization in dubbing: A translation approach," in *Topics in Audiovisual Translation*, Pilar Orero, Ed. 2004, pp. 35–52, John Benjamins B.V.

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

[7] 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*, Online, July 2020, pp. 257–264, Association for Computational Linguistics.

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

[9] 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.

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

[11] Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi, "MuST-C: a Multilingual Speech Translation Corpus," in *Proc. NAACL*, 2019, pp. 2012–2017.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[13] Mattia Di Gangi, Robert Enyedi, Alessndra Brusadin, and Marcello Federico, "Robust neural machine translation for clean and noisy speech translation," in *Proc. IWSLT*, 2019.

[14] Nishant Prateek, Mateusz Lajszczak, Roberto Barra-Chicote, Thomas Drugman, Jaime Lorenzo-Trueba, Thomas Merritt, Srikanth Ronanki, and Trevor Wood, "In other news: A bi-style text-to-speech model for synthesizing newscaster voice with limited data," in *Proc. NAACL*, 2019, pp. 205–213.

[15] Javier Latorre, Jakub Lachowicz, Jaime Lorenzo-Trueba, Thomas Merritt, Thomas Drugman, Srikanth Ronanki, and Klimkov Viacheslav, "Effect of data reduction on sequence-to-sequence neural TTS," in *Proc. ICASSP*, 2019, pp. 7075–7079.

[16] Jaime Lorenzo-Trueba, Thomas Drugman, Javier Latorre, Thomas Merritt, Bartosz Putrycz, Roberto Barra-Chicote, Alexis Moinet, and Vatsal Aggarwal, "Towards Achieving Robust Universal Neural Vocoding," in *Proc. Interspeech*, 2019, pp. 181–185.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. ICMAI*. Springer, 2015, pp. 234–241.

[18] Andreas Jansson, Eric Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde, "Singing voice separation with deep u-net convolutional networks," in *Proc. ISMIR*, 2017.

[19] Heiner Löllmann, Emre Yilmaz, Marco Jeub, and Peter Vary, "An improved algorithm for blind reverberation time estimation," in *Proc. IWAENC*, 2010, pp. 1–4.

[20] Emanuel AP Habets, "Room impulse response generator," Tech. Rep. 2.4, Technische Universiteit Eindhoven, 2006.

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

[22] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Interspeech 2017*. Aug. 2017, pp. 498–502, ISCA.

[23] M. Artetxe and H. Schwenk, "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond," *TACL*, 2019.

[24] Eric W. Noreen, *Computer-Intensive Methods for Testing Hypotheses: An Introduction*, John Wiley & Sons, 1989.

[25] Y. Yang, D. Cer, M. Guo A. Ahmad, J. Law, N. Constant, G. Abrego, S. Yuan nad C. Tar, Y. Sung, B. Strope, and R. Kurzweil, "Multilingual Universal Sentence Encoder for Semantic Retrieval," *ACL*, 2020.

[26] N. Reimers and I. Gurevych, "Sentence-Bert: Sentence Embeddings using Siamese Bert-networks.," *EMNLP*, 2019.

[27] F. Feng, Y. Tang, D. Cer, N. Arivazhagan, and W. Wang, "Language-Agnostic Bert Sentence Embedding," *arXiv preprint arXiv:2007.01852*, 2020.

[28] Douglas Bates, Martin Mchler, Ben Bolker, and Steve Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[29] Douglas Bates, Reinhold Kliegl, Shravan Vasishth, and Harald Baayen, "Parsimonious Mixed Models," *arXiv:1506.04967*, 2015.