Prosodic Alignment for off-screen automatic dubbing

Yogesh Virkar[1], *Marcello Federico*[1], *Robert Enyedi*[1], *Roberto Barra-Chicote*[2]

[1]AWS AI Labs
[2]Alexa AI

{yvvirkar|marcfede|renyedi|rchicote}@amazon.com

Abstract

The goal of automatic dubbing is to perform speech-to-speech translation while achieving audiovisual coherence. This entails *isochrony*, i.e., translating the original speech by also matching its prosodic structure into phrases and pauses, especially when the speaker's mouth is visible. In previous work, we introduced a *prosodic alignment* model to address isochrone or *on-screen* dubbing. In this work, we extend the prosodic alignment model to also address *off-screen* dubbing that requires less stringent synchronization constraints. We conduct experiments on four dubbing directions – English to French, Italian, German and Spanish – on a publicly available collection of TED Talks and on publicly available YouTube videos. Empirical results show that compared to our previous work the extended prosodic alignment model provides significantly better subjective viewing experience on videos in which on-screen and off-screen automatic dubbing is applied for sentences with speakers mouth visible and not visible, respectively.

Index Terms: speech translation, text-to-speech, automatic dubbing, off-screen dubbing

1. Introduction

Automatic Dubbing (AD) is an extension of speech-to-speech translation that replaces speech in a video with speech in a different language while preserving as much as possible the viewer experience. Speech translation [1, 2, 3, 4] consists of recognizing a speech utterance in the source language, performing translation, and optionally resynthesizing speech in the target language. Use cases for speech translation include video conferencing, live lectures, etc. in which close to real-time response is needed. In contrast, AD is used to aid the localization of audiovisual content, a highly complex workflow [5] usually managed during post-production by dubbing studios. High quality video dubbing usually involves speech synchronization at the utterance level (isochrony), lip movement level (phonetic synchrony) and body movement level (kinetic synchrony). In the past, most work on AD [6, 7, 8, 9] addressed isochrony, i.e., translating original speech by optimally matching its sequence of phrases and pauses. The idea is to first machine translate the source transcript by generating output with roughly the same duration [10, 11] –i.e. in terms of number of characters or syllables – of the input. Next, the translation is segmented into phrases and pauses of the same duration as that of the original phrases. This step is called prosodic alignment (PA).

Past work on PA [6, 7, 8, 9] focused on isochrony in the context of on-screen dubbing, i.e., dubbing of videos in which the speaker's mouth is visible for all utterances. However, in practical settings, it is quite common that videos contain scenes in which the speaker is not visible (off-screen) and for which the synchronization constraints of isochrone dubbing can be relaxed. Another example is that of automatic voiceover [12]. To address this case, we extend PA with a mechanism to address on/off-screen dubbing in which all or some of the sentences in a video are off-screen. We perform automatic and human evaluations that compare our original PA model for isochrone dubbing [9] with the augmented PA model for on/off dubbing [1]. We report results on a test set of TED talks extracted from the MUST-C corpus [13] and on 3 publicly available YouTube videos, on four dubbing directions, English (en) to French (fr), Italian (it), German (de) and Spanish (es). To summarize:

- We extend the PA model [9] to address off-screen dubbing.
- We introduce an automatic metric to compute intelligibility of dubbed videos.
- We run extensive automatic and subjective human evaluations comparing previous work with the new PA model on TED Talks and YouTube clips.
- Finally, we demonstrate the utility of automatic metrics in predicting human score by using linear mixed-effects models.

Our paper is organized as follows: First, we describe our dubbing architecture, then, we focus on existing and new PA methods and finally present and discuss experimental results comparing past and current work.

2. Dubbing Architecture

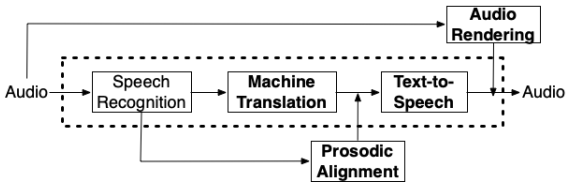

Figure 1: *Speech translation pipeline (dotted box) with enhancements introduced to perform automatic dubbing (bold).*

We build on the automatic dubbing architecture presented in [8, 7]. Figure 1 shows (in bold) how we extend a speech-to-speech translation [1, 2, 3] pipeline with: neural machine translation (MT) able to control verbosity of the output [11, 14, 15, 16, 17]; prosodic alignment (PA) [6, 8, 9] which addresses phrase-level synchronization of the MT output by leveraging the force-aligned source transcript; neural text-to-speech (TTS) [18, 19, 20] with precise duration control; and, finally, audio rendering that enriches TTS output with the original background noise (extracted via audio source separation [21, 22]) and reverberation, estimated from the original audio [23, 24].

[1]*For examples of dubbed videos with the latest PA model see* `https://github.com/amazon-research/on-off-screen-prosodic-alignment`.

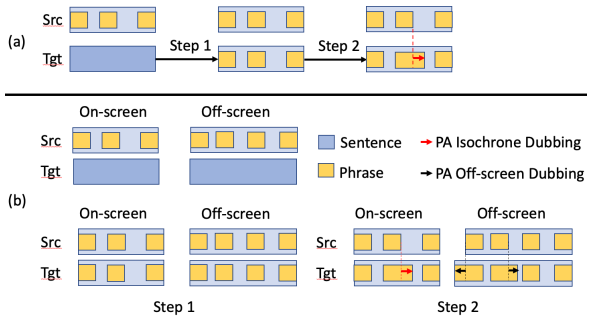

Figure 2: *Overview of dubbing conditions: (a) Isochrone dubbing [9] and (b) On/Off dubbing. The length of a box corresponds to the duration.*

3. Related Work

In the past there has been little work to address isochrony in dubbing [6, 7, 8, 9]. The approach of [6] involved generating and rescoring segmentation hypotheses by utilizing the attention weights of neural machine translation. While they focused only on the linguistic content matching between source-target phrases, Federico et al. [7] focused on fluency. In particular, their model utilized source-target duration matches and dynamic programming search for faster implementation. Their subsequent works [8, 9] further enhanced prosodic alignment (PA) by addition of features controlling for speaking rate variation and linguistic content matching. Additionally, they introduced time-boundary relaxation to further improve speaking rate control. However, none of these works focused on relaxing isochrony constraints by considering if the speaker is on-screen or off-screen. Recently, [25] leveraged on/off screen information to improve MT of dubbing scripts. Their rationale is that as human translations of scripts used in training reflect the different sync requirements posed by on-screen and off-screen speech, and hence it is worth introducing the same bias in the neural MT model. Our work complements [25], by showing how to leverage the same information in order to improve PA.

4. Prosodic Alignment

4.1. Isochrone Dubbing

PA aims to segment a translation to optimally match the sequence of phrases (or segments) and pauses of the corresponding source utterance. Let $\mathbf{e}$ denote the source sentence with n words and k breakpoints denoted by $\mathbf{i} = i_1, \ldots, i_k$ such that $1 \leq i_1 \leq i_2 \leq \ldots i_k = n$. Let T denote the temporal duration of $\mathbf{e}$ and let $\mathbf{s}$ denote a temporal segmentation into k segments where $\Delta\epsilon$ is the minimum silence after and before each break point.[2] Given the target sentence $\mathbf{f}$ of m words, the goal of PA is to find k breakpoints $\mathbf{j} = 1 \leq j_1 < j_2 < \ldots j_k = m$ within $\mathbf{f}$ that maximize the probability:

$$\max_{\mathbf{j}} \log \Pr\left(\mathbf{j}|\mathbf{i}, \mathbf{e}, \mathbf{f}, \mathbf{s}\right). \quad (1)$$

Assuming a Markovian model of $\mathbf{j}$, we get:

$$\log \Pr\left(\mathbf{j}|\mathbf{i}, \mathbf{e}, \mathbf{f}, \mathbf{s}\right) = \sum_{t=1}^{k} \log \Pr\left(j_t|j_{t-1}; t, \mathbf{i}, \mathbf{e}, \mathbf{f}, \mathbf{s}\right) \quad . \quad (2)$$

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

In [8] we derive a recurrent formular which permits to efficiently solve (2) with dynamic programming. Moreover, we allow target segments to possibly extend or contract the duration of the corresponding source interval by some fraction of $\Delta\epsilon$ to the left and to the right, denoted by δ_l and δ_r respectively, s.t., $\delta_l, \delta_r \in \left\{0, \pm\frac{1}{4}, \pm\frac{2}{4}, \pm\frac{3}{4}, \pm\frac{4}{4}\right\}$. In this way, we trade off strict isochrony for small adjustments to the speaking rate, in order to improve the viewing experience. In the past work [9] it was observed that relaxations on isochrony do not help improve the accuracy of finding optimal segmentation but only help improve speech fluency.

Two-step optimization procedure: For the above reasons, in [9], we introduce a two-step optimization procedure. In Step 1, we optimize the weights of the following log-linear model by maximizing segmentation accuracy over a manually annotated data set:

$$\log \Pr(j|j^{'}; t) \propto \sum_{a=1}^{4} w_a \log s_a(j, j^{'}; t) \quad (3)$$

The feature functions s_a of the model (notice that we dropped some of the dependencies in eq. (2) for readability) denote: (1) the language model score of target break point, (2) the cross-lingual semantic match score across source and target segments, (3) the speaking rate variation across target segments and (4) the speaking rate match across source and target segments respectively.

In Step 2, starting from given breakpoints $\hat{\mathbf{j}}$, we optimize the relaxations δ_l and δ_r for the t-th segment using another recurrent equation, that can also be solved via dynamic programming, derived from the log-linear model:

$$\log \Pr\left(\delta_l, \delta_r| \ldots; t\right) \propto \sum_{a=1}^{5} w_a \log s_a \left(\delta_l, \delta_r, \hat{j}_t, \hat{j}_{t-1}, \delta_l^{'}, \delta_r^{'}; t\right) \quad (4)$$

which includes as additional feature the isochrony score s_5 [9]. Weight w_5 is optimized by maximizing speech smoothness [9] over the training set, assuming the reference breakpoints $\hat{\mathbf{j}}$ are given. Speech smoothness measures speaking rate variations across contiguous segments. 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, [3], i.e:

$$r_e(t) = \frac{\text{duration}(\text{TTS}_e(\tilde{e}_t))}{|s_t|} \quad (5)$$

$$r_f(t) = \frac{\text{duration}(\text{TTS}_f(\tilde{f}_t))}{|s_t^*|} \quad (6)$$

4.2. On/Off Dubbing

Figure 2(a) shows that in our past work [9], during inference we apply the two steps of the PA component at the level of a single sentence, i.e., for each target sentence we first segment and then find the optimal relaxation using trained model defined by Eqs. (3), (4). In this work we focus on the more general use case

[3]We run TTS on the entire sentence, force-align audio with text [26, 27] and compute segment duration from the timestamps of the words. We can also compute duration using an explicit duration model [28].

of dubbing videos in which some of the sentences are on-screen, i.e., the speaker's mouth is visible, and some are off-screen, i.e., the speaker's mouth is not visible. We name this general case on/off dubbing. For off-screen sentences, we do not need the stringent requirement of isochrony and we can perform more relaxation to further improve the speaking rate control. Figure 2(b) gives an overview of the algorithm for on/off dubbing which extends the isochrone dubbing algorithm as follows:

Step 1 We apply the segmentation step (Eq. (3)) for all sentences, i.e., both on and off screen sentences.

Step 2 We apply the relaxation step locally for all on-screen sentences using Eq. (4) and globally across all off-screen sentences by replacing Eq. (4) with the following:

$$\Pr\left(\delta_l, \delta_r | \ldots; t\right) \propto \begin{cases} 1 & \text{if } r_f(t) \leq 1 \ , \\ 2 - r_f(t) & \text{if } 1 < r_f(t) \leq 2 \ , \\ 0 & \text{if } r_f(t) > 2 \end{cases} \tag{7}$$

In Step 2 we also apply a more relaxed policy in allocating relaxations δ_l and δ_r inside off-screen sentences. In particular, we allow using the entire available inter-phrase and inter-sentence intervals rather than limiting them to maximum $\Delta\epsilon$. Isochrone dubbing utilizes relaxation mechanism locally inside each sentence since for on-screen sentences, we tradeoff isochrony for improved speaking rate control. In contrast, for off-screen sentences we do not need isochrony and hence we utilize a *global relaxation mechanism* by computing optimal relaxations across all off-screen sentences using dynamic programming.

Regarding the scoring function (7), when the target speaking rate $r_f(t)$ is below 1, it returns a maximum score since for off-screen phrases we can contract time boundaries by setting the speaking rate $r_f(t)$ to 1 without consequences. When the speaking rate of a target phrase $r_f(t)$ is larger than 2, it returns the lowest score since too high speaking rates will result in low quality TTS speech. For $1 \leq r_f(t) \leq 2$, it returns scores that increasingly penalize values larger than 1, as they will correspond to less and less intelligible TTS speech.

5. Evaluation Data and Metrics

For training and evaluation, we re-translated and annotated video clips from 20 TED talks of the MUST-C corpus [13] and 3 YouTube videos by vloggers (see Sec. 8 of [29]). Each video clip contains 4 sentences manually annotated for on or off screen [4]. A single sentence contains one or more pauses of at least 300ms that are detected by force-aligning the source language English audio with text [26]. We manually collected and segmented translations in 4 target languages - French, Italian, German and Spanish - using external vendors to fit duration and segmentation of corresponding English utterances.

Overall, we created two test sets to test on/off dubbing PA (ON/OFF) against Isochrone dubbing PA (ISO), (i) D_1: 15 4-sentence clips in which all clips have all sentences being off-screen, (ii) D_2: 15 4-sentence clips in which all clips have at least one sentence being on-screen[5]. Compared to our previous work [9], we increase the size of extracted clips from 1 sentence to 4 sentences to test if the global relaxation mechanism provides better subjective viewing experience. To automatically estimate quality of dubbing, similar to [8, 9] we define Fluency (Fl) and Smoothness (Sm). For Smoothness we consider contiguous segments that span an entire 4-sentence video clip. We additionally introduce the following metric:

Intelligibility (In) of audio by dubbing target sentences $\mathbf{f}$ using prosodic alignment is defined by the ratio:

$$In(\mathbf{f}) = \frac{1 - \mathrm{WER}(\mathrm{TTS_f}(\mathrm{PA}(\mathbf{f})))}{1 - \mathrm{WER}(\mathrm{TTS_f}(\mathbf{f}))} \tag{8}$$

where WER is the word error rate by an automatic speech recognition system [6] run on TTS audio, either with prosody-alignment (numerator) or without prosody alignment (denominator).

			D_1		D_2	
			ISO	ON/OFF	ISO	ON/OFF
MuST-C	fr	Sm	68.5	75.3°	60.7	69.3*
		Fl	76.7	83.3	61.3	72.6
		In	93.6	93.5	93.5	93.2
	it	Sm	58.7	75.3*	52.0	68.3*
		Fl	68.3	80.0°	54.0	68.3*
		In	117.2	121.1*	98.8	99.0
	de	Sm	66.4	79.7*	57.6	70.4*
		Fl	81.7	86.7	58.6	74.1*
		In	94.3	94.3	91.7	93.0
	es	Sm	71.6	82.0*	61.9	76.0*
		Fl	80.0	85.0	61.9	79.4*
		In	124.9	125.7	97.0	98.0
YouTube	fr	Sm	70.6	80.9*	70.7	73.2
		Fl	66.7	80.0	60.0	60.0
		In	102.7	103.9	99.8	102.8
	it	Sm	73.6	81.3*	66.1	64.7
		Fl	40.0	73.3	46.7	43.8
		In	109.5	111.3	101.4	101.5
	de	Sm	69.9	82.3*	61.9	67.1°
		Fl	53.3	66.7	46.7	53.3
		In	102.7	105.6*	93.8	100.3°
	es	Sm	70.1	78.1*	67.0	72.0*
		Fl	33.3	60.0	60.0	66.7
		In	105.6	108.1	106.1	107.7°

Table 1: *Automatic evaluation of PA variants in terms of Smoothness (Sm), Fluency (Fl), Intelligibility (In) of: isochrone dubbing PA[9] (ISO), mixed dubbing PA (ON/OFF) applied on off-screen clips (D_1) or on mixed off-screen and on-screen clips (D_2). All test sets consist of 15 4-sentence video clips for each domain (MuST-C, YouTube). Significance testing is done with levels $p < 0.05$ ($^\circ$) and $p < 0.01$ (*).*

6. Experiments

6.1. Automatic Evaluation

Table 1 shows the results for automatic evaluation. We observe that ON/OFF outperforms ISO on MUST-C D_1, with respect to Sm(oothness) and Fl(uency) with relative improvements ranging from 9.9%-28.3% and 6.1%-17.1% respectively, while for In(telligibility) ON/OFF provides 0.6%-14.5% improvements for it, de, es. Similar improvements are obtained for ON/OFF

[4] We consider the mixed case, in which the speaker's mouth is visible only for some part of the sentence, to be on screen to ensure isochrony.

[5] Testing on D_1 (baseline) is done to ensure we obtain expected results. Both D_1, D_2 contain the same 15 clips across all languages.

[6] We use the off-the-shelf service Amazon Transcribe (https://aws.amazon.com/transcribe).

			D_1		D_2	
			ISO	ON/OFF	ISO	ON/OFF
MuST-C	fr	W	18.3	38*	21	43*
		S	4.43	4.79*	4.35	4.71*
	it	W	23	53.7*	15.3	54.3*
		S	4.61	5.36*	4.87	5.64*
	de	W	23.3	55.7*	19.7	64.7*
		S	4.45	5.11*	3.92	5.04*
	es	W	16.7	36.7*	28.7	37.3*
		S	5.03	5.35*	5.21	5.3
YouTube	fr	W	21.5	58.2 *	20.0	60.0*
		S	5.06	5.77*	4.68	5.35*
	it	W	17.3	68.7*	21.7	55.0*
		S	5.03	6.16*	5.08	5.87*
	de	W	20.3	56.7*	25.7	50.0*
		S	5.35	6.17*	5.00	5.53*
	es	W	27.0	56.7*	36.3	46.3*
		S	4.70	5.36*	4.95	5.09

Table 2: *Human evaluations using Wins (W) and Score (S) with prosodic alignments: (ISO) previous work on Isochrone dubbing [9], (ON/OFF) new PA model for dubbing applied on off-screen clips (D_1) or on mixed off-screen and on-screen clips (D_2). All test sets consist of 15 4-sentence video clips for each domain (MuST-C, YouTube). Significance testing is done with levels $p < 0.05$ ($^\circ$) and $p < 0.01$ (*).*

against ISO for MuST-C on D_2 and YouTube on D_1 and D_2. We note that improvements for D_1 are higher compared to D_2, primarily because D_1 considers videos in which all sentences are off-screen. This provides more opportunities for ON/OFF to exploit the global relaxation mechanism. Although, not all improvements are statistically significant, these metrics are well correlated with human judgements (see Sec. 6.2).

Compared to MuST-C, we find that YouTube data always obtains higher In scores. To investigate this, we computed the length compliance (LC) metric of [14] at the phrasal level that measures the percentage of translations whose length in characters is within $\pm 10\%$ of the length of the source. We found that on average, LC for YouTube was 19.8% higher than that for MuST-C. Higher value of LC allows us to better fit the translations in the available phrase intervals resulting in higher In.

6.2. Human Evaluation

In this section, we present results of human evaluation on the test set. For each dubbing direction and dataset we report results on 15 video clips extracted for each evaluation using the criterion noted in Sec. 5. We asked 20 native speakers to rate the subjective experience for viewing each dubbed video from each dubbing condition on a scale of 0-10. To reduce cognitive load, we compare two dubbing conditions for each evaluation and collect a total of 600 scores. For all dubbing conditions we utilize post-edited translations to focus the subjects on the synchronization aspect of dubbing.

Finally, for each evaluation we compare two conditions in a head-to-head manner and report Wins (percentage of times one condition is preferred over the other) and Score (average subjective score of dubbed videos) metrics. To measure the impact of PA model on human score, we use a linear-mixed-effects model (LMEM) [7] by defining subjects and clips as random effects [31].

[7] We used the lme4 package for R [30]

		MuST-C		YouTube	
		D_1	D_2	D_1	D_2
MuST-C	D_1	0.51*	0.43*	0.73*	0.70*
	D_2	0.07	0.26°	0.50*	0.50*
YouTube	D_1	0.14	0.33°	0.68*	0.65*
	D_2	0.43*	0.47*	0.68*	0.7*

Table 3: *Pearson correlation coefficient between predicted score from a LMEM model with fixed effects Sm, Fl, In and the averaged human score. We train a LMEM model on dataset in each row and predict score on dataset in each column. Significance testing is done with levels $p < 0.05$ ($^\circ$) and $p < 0.01$ (*).*

The results are summarized in Table 2. For the dubbing evaluations, we compare ON/OFF vs ISO on test sets D_1 and D_2 in two separate evaluations. ON/OFF clearly outperforms ISO on D_1 providing relative improvements in Wins on both MuST-C ranging from 107.7%-139.1% and YouTube ranging from 110%-297.1% with all results being statistically significant ($p < 0.01$). Similarly, ON/OFF outperforms ISO on D_2. Note that evaluations for D_1, D_2 are done by different subjects and hence the differences in Wins between D_1, D_2 are not comparable. Finally for Score, we obtain similar relative improvements on both MuST-C (1.7%-28.6%) and YouTube (2.8%-22.5%) with all improvements except the ones for es on D_2 being statistically significant ($p < 0.01$).

Relation between automatic and human scores: To explain the observed score variations using automatic metrics, we utilize LMEMs by aggregating evaluation data across all four languages. We define automatic metrics as fixed effects and subjects, clips, PA models and target languages as random effects. Our analysis reveals that Sm is the most impactful metric which is always statistically significant.

To further test if the learned LMEM can help predict human score, for every example we compute the average human score and compare it with the predicted score. We average out the random effect of subjects since each subject uses different score range. We train LMEM models, one on each dataset and domain, for a total of 4 models and predict scores on all 4 datasets. Table 3 shows that in most cases we obtain a statistically significant positive pearson correlation coefficient between the predicted and average human score. For MuST-C test sets we obtain small to medium correlation (< 0.5) while for YouTube test sets we obtain high correlation (≥ 0.5). Small correlation values cause inconsistency between magnitude of predicted and actual score differences. However it doesn't impact the sign of score differences. Hence, despite the small correlation, each model is able to correctly predict on average the winning system (ISO or On/Off) for all datasets.

7. Conclusions

We extended prosodic alignment to address off-screen dubbing that requires less stringent synchronization constraints by introducing a global relaxation algorithm. Using dynamic programming, this algorithm allows us to relax timing constraints across all off-screen sentences. Both automatic and human evaluations show that compared to applying isochrone dubbing for all sentences, relaxing the synchronization constraints for off-screen sentences significantly improves model performance on both automatic and subjective metrics. Finally, using the linear mixed-effects models we show that a linear combination of all automatic metrics correlates well with the average human score.

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] 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.

[3] L. C. 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.

[4] M. Sperber and M. 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*, 2020, pp. 7409–7421.

[5] F. Chaume, "Synchronization in dubbing: A translation approach," in *Topics in Audiovisual Translation*, pp. 35–52. 2004.

[6] A. Öktem, M. Farrús, and A. Bonafonte, "Prosodic Phrase Alignment for Machine Dubbing," in *Proceedings of Interspeech*, Graz, Austria, 2019, arXiv: 1908.07226.

[7] 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 *Proceedings of the 17th International Conference on Spoken Language Translation*, Online, July 2020, pp. 257–264, Association for Computational Linguistics.

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

[9] Y. Virkar, M. Federico, R. Enyedi, and R. Barra-Chicote, "Improvements to Prosodic Alignment for Automatic Dubbing," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2021, pp. 7543–7574, ISSN: 2379-190X.

[10] A. Saboo and T. 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] S. M. Lakew, M. Di Gangi, and M. Federico, "Controlling the Output Length of Neural Machine Translation," in *Proceedings of IWSLT*, Hong Kong, China, Oct. 2019, arXiv: 1910.10408.

[12] Mattia Di Gangi, Nick Rossenbach, Alejandro Pérez, Parnia Bahar, Eugen Beck, Patrick Wilken, and Evgeny Matusov, "Automatic video dubbing at AppTek," in *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, Ghent, Belgium, June 2022, pp. 349–350, European Association for Machine Translation.

[13] 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.

[14] Surafel M. Lakew, Yogesh Virkar, Prashant Mathur, and Marcello Federico, "Isometric mt: Neural machine translation for automatic dubbing," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6242–6246.

[15] Patrick Wilken and Evgeny Matusov, "AppTek's submission to the IWSLT 2022 isometric spoken language translation task," in *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, Dublin, Ireland (in-person and online), May 2022, pp. 369–378, Association for Computational Linguistics.

[16] Zongyao Li, Jiaxin Guo, Daimeng Wei, Hengchao Shang, Minghan Wang, Ting Zhu, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Lizhi Lei, Hao Yang, and Ying Qin, "HW-TSC's participation in the IWSLT 2022 isometric spoken language translation," in *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, Dublin, Ireland (in-person and online), May 2022, pp. 361–368, Association for Computational Linguistics.

[17] Aakash Bhatnagar, Nidhir Bhavsar, Muskaan Singh, and Petr Motlicek, "Hierarchical multi-task learning framework for isometric-speech language translation," in *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, Dublin, Ireland (in-person and online), May 2022, pp. 379–385, Association for Computational Linguistics.

[18] 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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, Minneapolis, Minnesota, 2019, pp. 205–213, Association for Computational Linguistics.

[19] 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," *arXiv:1811.06315 [cs, eess]*, 2018, arXiv: 1811.06315.

[20] 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*, 2019.

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

[22] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-net convolutional networks," in *Proceedings of the 18th International Society for Music Information Retrieval Conference*, Suzhou, China, 2017, p. 8.

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

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

[25] A. Karakanta, S. Bhattacharya, S. Nayak, T. Baumann, M. Negri, and M. Turchi, "The Two Shades of Dubbing in Neural Machine Translation," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online), 2020, pp. 4327–4333, International Committee on Computational Linguistics.

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

[27] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Interspeech*, 2017, pp. 498–502.

[28] Johanes Effendi, Yogesh Virkar, Roberto Barra-Chicote, and Marcello Federico, "Duration modeling of neural tts for automatic dubbing," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8037–8041.

[29] Antonios Anastasopoulos et al, "Findings of the IWSLT 2022 evaluation campaign," in *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, Dublin, Ireland (in-person and online), May 2022, pp. 98–157, Association for Computational Linguistics.

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

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