Evaluating and Optimizing Prosodic Alignment for Automatic Dubbing

Marcello Federico Yogesh Virkar Robert Enyedi Roberto Barra-Chicote

Amazon

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

Abstract

Automatic dubbing aims at replacing all speech contained in a video with speech in a different language, so that the result sounds and looks as natural as the original. Hence, in addition to conveying the same content of an original utterance (which is the typical objective of speech translation), dubbed speech should ideally also match its duration, the lip movements and gestures in the video, timbre, emotion and prosody of the speaker, and finally background noise and reverberation of the environment. In this paper, after describing our dubbing architecture, we focus on recent progress on the prosodic alignment component, which aims at synchronizing the translated transcript with the original utterances. We present empirical results for English-to-Italian dubbing on a publicly available collection of TED Talks. Our new prosodic alignment model, which allows for small relaxations in synchronicity, shows to significantly improve both prosodic alignment accuracy and overall subjective dubbing quality of previous work.

Index Terms: speech translation, text to speech, automatic dubbing.

1. Introduction

Professional dubbing [1] is a complex and labor intensive process that involves many steps, roughly summarized by: extracting speech segments from the audio track and annotating these with speaker information; transcribing the speech segments; translating the transcript in the target language; adapting the translation for synchronization; casting the voice talents; performing the dubbing sessions; fine-aligning the dubbed speech segments; and, finally, mixing the new voice tracks within the original soundtrack. Adapting a translation for dubbing considers three types of synchronization [2]: (i) phonetic synchrony, which consists in adapting the translation to the articulatory movements of the performer on the screen, (ii) kinetic synchrony, which consists in producing a translation that is temporally consistent with the performer's body movements, (iii) and *isochrony*, the most critical form synchronization of dubbing[1] and the focus of our work, which consist in arranging the translation into phrases so that the voice talent can closely match the start and end points of the performer's speech activity, while preserving fluency and the original speaking rate.

This paper builds on the automatic dubbing architecture presented in [3] (Figure 1) that extends a speech-to-speech translation [4, 5, 6] pipeline [2] with: neural machine translation (MT) robust to ASR errors and able to control verbosity of the output [9, 10, 11]; prosodic alignment (PA) [12] which addresses isochrony by leveraging temporally segmented ASR output; neural text-to-speech (TTS) [13, 14, 15] with precise duration control; audio rendering that enriches TTS output with the original background noise (extracted via audio source separation with deep U-Nets [16, 17]) and reverberation, estimated from the original audio [18, 19]. In [3], we also run a subjective evaluation of the naturalness of 25 video clips of TED Talks, extracted from the MUST-C [20] corpus, automatically dubbed from English to Italian. Results confirmed what already reported in [12]: perceived dubbing quality significantly drops when PA splits the translation into phrases that, due to the required time boundaries, result in TTS speech that is either too slow, too fast, or too uneven across consecutive phrases.

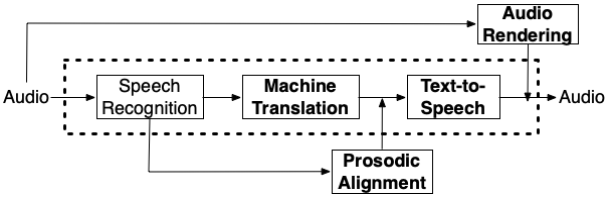

Figure 1: *Speech-to-speech translation pipeline (dotted box) with enhancements to perform automatic dubbing (in bold).*

In this paper we discuss an improved version of the PA model presented in [3] that, while segmenting the translation to optimally match the speaking rate of the original phrases also allows for small and tolerable relaxations of isochrony[2] for the sake of avoiding unnatural TTS output. In the following sections, we provide background knowledge about automatic dubbing, introduce our new PA model, present a speech dubbing data set and objective metrics to evaluate PA, and present and discuss results of an automatic and a subjective evaluation of our new approach.

2. Background

There is relatively little work on automatic speech dubbing and mostly tackling the problem as replacing natural speech with synthetic speech in the same language. In [21], natural and synthetic speech are time-aligned at the frame level with a dynamic time warping algorithm. Similarly, in [22, 23] speech generated from subtitles is aligned to the original audio track by shortening the length of the script, with fine-grained control of the duration of TTS, and relaxation of the timing constraints given by the subtitles. A completely different approach to address cross-lingual phonetic synchrony is proposed in [24] by directly manipulating the video showing the actor's mouth. The work that pioneered cross-lingual synchronization at the phrase/utterance level (isochrony) is [12], which introduces the concept of prosodic alignment. Starting from a transcript of the original audio, annotated with temporal phrase-boundaries, and output from a neural MT model, the PA model aligns the translation to the source phrases by leveraging the attention weights of the neural MT model. More recently, we proposed a PA

[1] Poorly dubbed movies are often grounded in poor isochrony [2].

[2] While end-to-end neural models look appealing and versatile [7, 8], we choose a modular approach to also explore the integration of automatic and human dubbing.

source:	"He asked Octavio"
start-end:	0.78s - 1.35s
target:	"Chiese a Octavio"
source:	"to be his chief of staff."
start-end:	1.87s - 3.24s
target:	"di fargli da capo del personale."

Table 1: *Example of prosodic alignment.*

model [3] which does not require cross-lingual information, but guides the search for the optimal alignment with two types of information: the speaking rate match between corresponding source-target phrases and the linguistic plausibility of the chosen split points. In particular, the speaking rate match is directly computed at the string level and the linguistic plausibility of a break point (pause) is evaluated with a language model.

3. Prosodic Alignment

Given the source sentence/utterance: *"He asked Octavio to be his chief of staff"*, temporally segmented as in Table 1, and its Italian translation *"Chiese a Octavio di fargli da capo del personale"*, the goal of prosodic alignment is to segment the target sentence in a way to optimally match the sequence of phrases and pauses of the source sentence. Let $\mathbf{e} = e_1, e_2, \ldots, e_n$ be a source sentence of n words, segmented according to k breakpoints $1 \leq i_1 < i_2 < \ldots i_k = n$, shortly denoted with $\mathbf{i}$. Let the temporal duration of $\mathbf{e}$ be T and the temporal intervals of the segmentation $\mathbf{i}$ be $s_1 = [l_1, r_1], \ldots, s_k = [l_k, r_k]$, shortly 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.[3] Given a target sentence $\mathbf{f} = f_1, f_2, \ldots, f_m$ of m words, the goal is to find k breakpoints $1 \leq j_1 < j_2 < \ldots j_k = m$ (shortly 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. However, this model implicitly assumes that corresponding source and target segments, defined by $\mathbf{i}$ and $\mathbf{j}$, have exactly the same duration (isochrony), defined by $\mathbf{s}$. We instead relax this constraint by actually allowing 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 such that $\delta_l, \delta_r \in \left\{0, \frac{1}{4}, \frac{2}{4}, \frac{3}{4}, \frac{4}{4}\right\}$. Hence, the idea is to extend the search to solutions that can possibly relax the isochrony constraint. Thus, we optimize,

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

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

$\mathbf{e}, \mathbf{f}$	source and target word sequences
i, j	word positions in $\mathbf{e}$ and $\mathbf{f}$
k	number of breakpoints in $\mathbf{e}$
$\mathbf{i}$,$\mathbf{j}$	sequences of k breakpoints in $\mathbf{e}$ and $\mathbf{f}$
t	index over segments in $\mathbf{e}$ and $\mathbf{f}$
$\tilde{e}_t$, $\tilde{f}_t$	source and target phrases of t-th segment
s_t	original temporal interval of t-th source segment
s_t^*	relaxed temporal interval of t-th target segment
$r_e(t)$	speaking rate of source phrase $\tilde{e}_t$ in interval s_t
$r_f(t)$	speaking rate of target phrase $\tilde{f}_t$ in interval s_t^*
δ_l, δ_r	left and right relaxations in the range $0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$

Table 2: *Notation*

Where now 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 $t-1$.
We define the model probability in (4) with a log-linear model:

$$\log \Pr\left(j, \delta_l, \delta_r \mid \ldots; t\right) \propto \sum_{k=1}^{4} w_a \log s_a\left(j, \delta_l, \delta_r, \ldots; t\right) \tag{5}$$

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

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

Speaking rate computations rely on the strings $\tilde{f}_t$ and $\tilde{e}_t$, composing 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,[4] i.e:

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

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

3.1. Speaking rate variation

By abusing the notation in (5), we define the speaking-rate variation $s_{sv}(\cdot)$ between consecutive target segments by:

$$s_{sv}\left(\tilde{f}_t, s_t^*, \tilde{f}_{t-1}, s_{t-1}^*; t\right) = 1 - \frac{|r_f(t) - r_f(t-1)|}{r_f(t-1)} \quad . \tag{8}$$

The objective of this feature is to penalize sentence split that generate TTS speech with speaking rate varying too much across consecutive segments. This feature reaches the maximum when consecutive segments have the same speaking rate and is activated starting from the second segment ($t \geq 2$).

[4] We run TTS on the entire sentence, force-align audio with text [25, 26] and compute segment duration from the time-stamps of the words.

3.2. Speaking rate match

We define the speaking-rate match $s_{sm}(\cdot)$ between corresponding source and target segments by:

$$s_{sm}\left(\tilde{e}_t, \tilde{f}_t, s_t, s_t^*; t\right) = 1 - \frac{|r_f(t) - r_e(t)|}{r_e(t)} \tag{9}$$

The rationale of this feature is to favor sentence splits that generate TTS utterances that closely track the speaking rate of the original audio. This feature reaches its maximum when the target and source speaking rates are identical. It complements the previous feature by (i) rewarding speaking rate variations in the target that are also present in the source and (ii) penalizing lack of variations in the target speaking rate when such variations are indeed present in the source. Empirically, we found beneficial regularizing $r_e(t)$ by clipping it to the range 60%-140%.

3.3. Isochrony score

We define the isochrony score $s_{is}(\cdot)$ for relaxations δ_l, δ_r, as:

$$s_{is}(\delta_l, \delta_r) = 1 - \left[\alpha\, \delta_l + (1-\alpha)\, \delta_r\right] \tag{10}$$

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. On the other hand, when some relaxation is useful to reduce the TTS speaking rate, the score penalizes left boundary relaxations more that right boundary relaxations. In fact, we empirically observed that small relaxations of isochrony in dubbed videos are more tolerable after the actress stops speaking than before the actress starts speaking. (Our observation is also confirmed in [2]). We thus set α, so that left relaxation is always penalized more than right relaxation, i.e.:

$$\alpha\delta_l > (1-\alpha)\delta_r \quad \forall \delta_r, \delta_l \in \{0, \frac{1}{4}, \frac{2}{4}, \frac{3}{4}, 1\} \tag{11}$$

which is satisfied by $\alpha > \frac{4}{5}$.

3.4. Language model

The language model score $s_lm(\cdot)$ estimates the probability of a break between the target strings $\tilde{f}_{t-1}$ and $\tilde{f}_t$ [3] by:

$$\begin{aligned} s_{lm}(j, \tilde{f}_{t-1}, \tilde{f}_t) &= \Pr\left(br | \tilde{f}_{t-1}, \tilde{f}_t\right) \\ &= \frac{p\left(\tilde{f}_{t-1}, br, \tilde{f}_t\right)}{p\left(\tilde{f}_{t-1}, br, \tilde{f}_t\right) + p\left(\tilde{f}_{t-1}, \tilde{f}_t\right)} \end{aligned} \tag{12}$$

We found convenient mapping words to part-of-speech, leveraging a part-of-speech 3-gram language model (details in Section 5) and to map the break symbol br to a punctuation-class denoting a *pause*, which includes period, comma, column and semicolon. This simple feature, which is also used in our previous PA model [3], represents a reasonable starting point while investigating more advanced prediction models [27, 28].

4. Evaluation Data and Metrics

For training and evaluation purposes we extracted and annotated 120 video clips from 20 TED talks of the MUST-C test set, containing exactly one sentence with one or more pauses of at least $300ms$. [5] For each clip we time aligned the English transcript with the audio and manually adapted and segmented the available Italian translations to fit the duration and segmentation of the corresponding English utterances (see example in Table 1). This process required more than one iteration, each ending with a dubbing step to assess the quality of the manually generated PA. The annotation process resulted in a total of 120 sentence pairs containing at least one break point, for a total of 187 breakpoints and 307 segments. Hence, we defined the following PA quality metrics for a set of sentences:

Accuracy is the percentage of sentences for which the PA segmentation coincides with the manual reference. We implicitly assume that missing one correct break point inside a sentence with multiple breakpoints does compromise the PA quality of the entire sentence.

Fluency is the percentage of sentences having TTS speaking rate for all segments in the range 60%-140%. We expect that segments with speaking rate outside this range might result in low fluency.

Smoothness measures the stability of the TTS speaking rate across all contiguous target segments:

$$\left\langle 1 - \frac{|r_f(t) - r_f(t-1)|}{r_f(t-1)} \right\rangle \ , \tag{13}$$

where $\langle . \rangle$ denotes the average over all segment pairs.

5. Experiments

Both our new and previous [3] PA model require estimating a part-of-speech language model (LM) feature as well as weights for all their features. We mapped target language word sequences into part-of-speech sequences with an online service[6], and estimated a 3-gram POS LM on the training portion of the MUST-C corpus [20] with the KENLM toolkit [29]. Regarding feature weights, the model in [3], in short model A, requires estimating weights for a speaking rate matching feature s_{sr} (based on character counts), and a language model feature s_{lm}, identical to the one discussed above. For our new PA model, we are interested in performing experiments without and with relaxation (isochrony feature), which we call model B and model C. We optimize feature weights of each PA model over accuracy by applying a hierarchical grid search [7] with the following nested convex combinations of feature pairs: [8]

$$\begin{aligned} &\text{A:} \quad (s_{lm}, s_{sm})_{w_{lm}} \\ &\text{B:} \quad (s_{lm}, (s_{sm}, s_{sv})_{w_{sm}})_{w_{lm}} \\ &\text{C:} \quad \left(s_{is}, (s_{lm}, (s_{sm}, s_{sv})_{w_{sm}})_{w_{lm}}\right)_{w_{is}} \end{aligned}$$

Note that for model C, we start from the weights of B, and set the weight of the isochrony feature to the minimum value providing the highest Accuracy, so to maximize its use.

In order to make optimal use of our 120 annotated sentences, we apply a 5-fold cross-validation scheme [30] for estimating and evaluating each PA model. Hence, we partition the data into 5 sets of 24 sentences by applying stratified sampling on the number of contained breakpoints.

[5] Detected by force-aligning original English audio with text [25].

[6] `https://aws.amazon.com/comprehend`.

[7] To simplify search and run initial ablation tests across models.

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

Automatic	A	B	C	R
Accuracy	49.17%	65.83% *	71.67% **	100%
Fluency	54.17%	71.67% **	89.17% **	68.33%
Smoothness	65.74%	81.33% **	87.40% **	73.15%

Manual	A vs. C		C vs. R	
Wins	28.5%	45.6% **	29.3%	36.9% **
Score	4.72	5.10 **	5.20	5.39 *

Table 3: *Automatic and manual evaluations with three prosodic alignments: (A) previous work [3], (B) new model without relaxation, (C) new model with relaxation, (R) manual reference. Test sets are made of 120 sentences and 50 video clips, respectively. Significance testing on automatic metrics is against model A, with levels* $p < 0.05$ *(*) and* $p < 0.01$ *(**).*

5.1. Automatic Evaluation

For each PA model and evaluation metric we report the average value on the 120 sentences. Results are reported in the upper section of Table 3. For models B and C we report statistical significance against model A by applying randomized paired permutation tests [31]. We observe that model B improves significantly over the baseline model A for all metrics (+33.9% relative in Accuracy, +32.3% in Fluency, +23.7% in Smoothness) confirming that even w/o relaxation the new scoring function already improves the previous model. Gains in Fluency and Smoothness seem to prove that model B can compute more accurate source-target duration matches and directly control speaking rate variations in the scoring function.

Comparison of A with the full model with relaxation (C), shows further gains across the board: +45.8% relative in Accuracy, +64.6% in Fluency and +32.9% in Smoothness, all statistically significant ($p < 0.01$). Further improvements in Fluency and Smoothness can be attributed to the ability of model C to lower too high speaking rates through the relaxation mechanism. Gains in Accuracy over model B are also due to the increased flexibility of model C in determining the optimal breakpoints, thanks to the relaxation mechanism. Finally, it is worth noticing that the reference segmentation (R) generates PA with lower Smoothness and Fluency than models B and C. We will discuss this aspect in the next subsection.

5.2. Manual Evaluation

We run a manual evaluation on a subset of 50 test sentences, selected among those with the highest number of breakpoints and such that A and C generate different PA (segmentation and time intervals). For each sentence and the corresponding video clip, we generated dubbed videos with the architecture described in the Introduction, by applying PA with A, C and R, followed by neural text-to-speech, background noise re-insertion and re-reverberation. We asked Italian speaking subjects to blindly grade their viewing experience of each dubbed video on a scale from 0 to 10. To make the cognitive load of the task acceptable, we run two distinct experiments, each comparing video clips dubbed under two conditions: A vs. C and C vs. R. The two evaluations were run each with 20 subjects from Amazon Mechanical Turk, and collected a total of $2,000$ scores.

For each experiment, we measured the percentage of times one condition was preferred over the other (Wins) together with its statistical significance [31]. We also measured the effect of each PA model on human scores using a *linear mixed-effects model*[9] (LMEM), by defining subjects and sentences as random effects [33]. Results are summarized in Table 3.

In the first experiment, model C clearly outperforms model A both in terms of Wins (+60% relative, $p < 0.01$) and Score (+8% relative, $p < 0.01$). The second experiment, comparing C vs. R shows smaller differences in Wins (−10.6% relative, $p < 0.01$) and Score (−3.5% relative, $p < 0.05$).[10]

Interestingly, model A and the reference (R) generate a segmentation by just looking at the source and target strings, but clearly humans are better at using this information to predict the duration of phrases and at guessing where to insert pauses. Model C, reduces the gap of A with more accurate predictions and with some tolerance in the time boundaries, which actually contributes both to increase the overall accuracy and the smoothness of the speaking rate. For model C, we also measured with LMEM the effect of relaxation on human Score. As relaxation is introduced at the segment level and sentences might include multiple segments, we consider both the maximum and total amount of relaxation introduced at the single sentence. For both effects, no statistical significant impact on Score was measured ($p < 0.225$ and $p < 0.073$, respectively).

By comparing manual and automatic scores in Table 3, it is clear that out of the three automatic metrics, Accuracy is the most relevant for final dubbing quality. We evaluated with LMEMs the impact of each automatic score on the human score with all data points, using subjects, sentences and systems as random effects. We found that only Accuracy has a statistically significant impact ($p < 0.001$). Notice that Accuracy of A and C on the manual evaluation are respectively of 30% and 83%. If we restrict the analysis on data of system C, we found that both Accuracy and Smoothness have statistically significant impact ($p < 0.001$), while Fluency has not. Our intuition is that our current definition of Fluency (=speaking rate falls in the range $[0.6, 1.4]$) should be probably narrowed or defined on a continuous range.

Finally, though not the main goal of this study, from comments left by our subjects, we ascertain that text-to-speech voice quality is an important factors that needs further improvement, for instance in the case of slow speaking rate or when the speaker makes many pauses. This explains why the scores of system R are on average low.

6. Conclusion

We presented and evaluated a new prosodic alignment model for automatic speech dubbing. The new model, in addition of segmenting a target language sentence to fit the timing of the corresponding multiple utterances in the original audio, includes some small relaxation of the time boundaries. We empirically observe that the addition of relaxations permits to significantly increase the accuracy of the segmentation as well as the smoothness and fluency of the generated text-to-speech. Manual evaluations, show that segmentation accuracy is the primary factor for quality on which we should focus on. Future work will be devoted to improve the quality and realism of text-to-speech, especially when specific speaking styles have to be mimicked.

[9] We used the `lme4` package for R [32].

[10] The difference in Score of C across the two experiments is likely due to the absence of a common reference anchor [34].

7. References

[1] X. Martínez, "Film dubbing, its process and translation," in *Topics in Audiovisual Translation*, P. Orero, Ed. John Benjamins B.V., 2004, pp. 3–8.

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

[3] M. Federico, R. Enyedi, R. Barra-Chicote, R. Giri, U. Isik, A. Krishnaswamy, and H. Sawaf, "From speech-to-speech translation to automatic dubbing," *arXiv:2001.06785*, 2020.

[4] 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.

[5] 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*. ISCA, Aug. 2017, pp. 2625–2629. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0503.html

[6] L. Cross Vila, C. Escolano, J. A. R. Fonollosa, and M. R. Costa-Juss, "End-to-End Speech Translation with the Transformer," in *IberSPEECH 2018*. ISCA, Nov. 2018, pp. 60–63. [Online]. Available: http://www.isca-speech.org/archive/IberSPEECH_2018/abstracts/IberS18_P1-9_Cross-Vila.html

[7] M. A. Di Gangi, V.-N. Nguyen, M. Negri, and M. Turchi, "Instance-Based Model Adaptation For Direct Speech Translation," in *ICASSP*, Oct. 2019, arXiv: 1910.10663. [Online]. Available: http://arxiv.org/abs/1910.10663

[8] M. A. Di Gangi, M. Negri, and M. Turchi, "One-To-Many Multilingual End-to-end Speech Translation," in *IEEE ASRU*, Oct. 2019, arXiv: 1910.03320. [Online]. Available: http://arxiv.org/abs/1910.03320

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

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

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

[12] A. Öktem, M. Farrùs, and A. Bonafonte, "Prosodic Phrase Alignment for Machine Dubbing," in *Proc. Interspeech*, 2019. [Online]. Available: http://arxiv.org/abs/1908.07226

[13] 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*, 2019, pp. 205–213.

[14] 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*, 2019, pp. 7075–7079.

[15] 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, pp. 181–185. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1424

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

[17] 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.

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

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

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

[21] W. Verhelst, "Automatic Post-Synchronization of Speech Utterances," in *Proc. Eurospeech*, 1997, pp. 899–902. [Online]. Available: https://www.isca-speech.org/archive/eurospeech_1997/e97_0899.html

[22] Z. Hanzlìček, J. Matoušek, and D. Tihelka, "Towards automatic audio track generation for Czech TV broadcasting: Initial experiments with subtitles-to-speech synthesis," in *Proc. Int. Conf. on Signal Processing*, 2008, pp. 2721–2724.

[23] J. Matoušek, Z. Hanzlìček, D. Tihelka, and M. Mèner, "Automatic dubbing of TV programmes for the hearing impaired," in *Proc. IEEE Signal Processing*, 2010, pp. 589–592.

[24] S. Furukawa, T. Kato, P. Savkin, and S. Morishima, "Video reshuffling: automatic video dubbing without prior knowledge," in *Proc. ACM SIGGRAPH*, 2016. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2945078.2945097

[25] R. M. Ochshorn and M. Hawkins, "Gentle Forced Aligner," 2017. [Online]. Available: https://lowerquality.com/gentle/

[26] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Interspeech 2017*. ISCA, Aug. 2017, pp. 498–502. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/1386.html

[27] O. Tilk and T. Alumae, "LSTM for Punctuation Restoration in Speech Transcripts," in *Interspeech*, 2015, pp. 683–687.

[28] O. Tilk and T. Alume, "Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration," in *Interspeech*, Sep. 2016, pp. 3047–3051. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2016/abstracts/1517.html

[29] K. Heafield, "KenLM: Faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 2011, pp. 187–197.

[30] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Science & Business Media, 2009.

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

[32] D. Bates, M. Mchler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015. [Online]. Available: https://www.jstatsoft.org/index.php/jss/article/view/v067i01

[33] D. Bates, R. Kliegl, S. Vasishth, and H. Baayen, "Parsimonious Mixed Models," *arXiv:1506.04967*, 2015. [Online]. Available: http://arxiv.org/abs/1506.04967

[34] *Method for the subjective assessment of intermediate quality level of coding systems*, International Communication Union, 2014, recommendation ITU-R BS.1534-2. [Online]. Available: "https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1534-1-200301-S!!PDF-E.pdf"