From Speech-to-Speech Translation to Automatic Dubbing

Marcello Federico **Robert Enyedi** **Roberto Barra-Chicote** **Ritwik Giri**

Umut Isik **Arvindh Krishnaswamy** **Hassan Sawaf***

Amazon

Abstract

We present enhancements to a speech-to-speech translation pipeline in order to perform automatic dubbing. Our architecture features neural machine translation generating output of preferred length, prosodic alignment of the translation with the original speech segments, neural text-to-speech with fine tuning of the duration of each utterance, and, finally, audio rendering to enriches text-to-speech output with background noise and reverberation extracted from the original audio. We report and discuss results of a first subjective evaluation of automatic dubbing of excerpts of TED Talks from English into Italian, which measures the perceived naturalness of automatic dubbing and the relative importance of each proposed enhancement.

1 Introduction

Automatic dubbing can be regarded as an extension of the speech-to-speech translation (STST) task (Wahlster, 2013), which is generally seen as the combination of three sub-tasks: (i) transcribing speech to text in a source language (ASR), (ii) translating text from a source to a target language (MT) and (iii) generating speech from text in a target language (TTS). Independently from the implementation approach (Weiss et al., 2017; Waibel, 1996; Vidal, 1997; Metze et al., 2002; Nakamura et al., 2006; Casacuberta et al., 2008), the main goal of STST is producing an output that reflects the linguistic content of the original sentence. On the other hand, automatic dubbing aims to replace all speech contained in a video document 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 the original utterance, dubbing should also match the original timbre, emotion, duration, prosody, background noise, and reverberation.

While STST has been addressed for long time and by several research labs (Waibel, 1996; Vidal, 1997; Metze et al., 2002; Nakamura et al., 2006; Wahlster, 2013), relatively less and more sparse efforts have been devoted to automatic dubbing (Matoušek et al., 2010; Matoušek and Vít, 2012; Furukawa et al., 2016; Öktem et al., 2019), although the potential demand of such technology could be huge. In fact, multimedia content created and put online has been growing at exponential rate, in the last decade, while availability and cost of human skills for subtitling and dubbing still remains a barrier for its diffusion worldwide.[1] Professional dubbing (Martínez, 2004) of a video file is a very labor intensive process that involves many steps: (i) extracting speech segments from the audio track and annotating these with speaker information; (ii) transcribing the speech segments, (iii) translating the transcript in the target language, (iv) adapting the translation for timing, (v) casting the voice talents, (vi) performing the dubbing sessions, (vii) fine-aligning the dubbed speech segments, (viii) mixing the new voice tracks within the original soundtrack.

Automatic dubbing has been addressed both in monolingual cross-lingual settings. In (Verhelst, 1997), synchronization of two speech signals with the same content was tackled with time-alignment via dynamic time warping. In (Hanzlìček et al., 2008) automatic monolingual dubbing for TV users with special needs was generated from subtitles. However, due to the poor correlation between length and timing of the subtitles, TTS output fre-

* Contribution while the author was with Amazon.

[1] Actually, there is still a divide between countries/languages where either subtitling or dubbing are the preferred translation modes (Kilborn, 1993; Koolstra et al., 2002). The reasons for this are mainly economical and historical (Danan, 1991).

quently broke the timing boundaries. To avoid unnatural time compression of TTS's voice when fitting its duration to the duration of the original speech, (Matoušek et al., 2010) proposed phone-dependent time compression and text simplification to shorten the subtitles, while (Matoušek and Vít, 2012) leveraged scene-change detection to relax the subtitle time boundaries. Regarding cross-lingual dubbing, lip movements synchronization was tackled in (Furukawa et al., 2016) by directly modifying the actor's mouth motion via shuffling of the actor's video frames. While the method does not use any prior linguistic or phonetic knowledge, it has been only demonstrated on very simple and controlled conditions. Finally, mostly related to our contribution is (Öktem et al., 2019), which discusses speech synchronization at the phrase level (prosodic alignment) for English-to-Spanish automatic dubbing.

In this paper we present research work to enhance a STST pipeline in order to comply with the timing and rendering requirements posed by cross-lingual automatic dubbing of TED Talk videos. Similarly to (Matoušek et al., 2010), we also shorten the TTS script by directly modifying the MT engine rather than via text simplification. As in (Öktem et al., 2019), we synchronize phrases across languages, but follow a fluency-based rather than content-based criterion and replace generation and rescoring of hypotheses in (Öktem et al., 2019) with a more efficient dynamic programming solution. Moreover, we extend (Öktem et al., 2019) by enhancing neural MT and neural TTS to improve speech synchronization, and by performing audio rendering on the dubbed speech to make it sound more real inside the video.

In the following sections, we introduce the overall architecture (Section 2) and the proposed enhancements (Sections 3-6). Then, we present results (Section 7) of experiments evaluating the naturalness of automatic dubbing of TED Talk clips from English into Italian. To our knowledge, this is the first work on automatic dubbing that integrates enhanced deep learning models for MT, TTS and audio rendering, and evaluates them on real-world videos.

2 Automatic Dubbing

With some approximation, we consider here automatic dubbing of the audio track of a video as the task of STST, i.e. ASR + MT + TTS, with the additional requirement that the output must be temporally, prosodically and acoustically close to the original audio. We investigate an architecture (see Figure 1) that enhances the STST pipeline with (i) enhanced MT able to generate translations of variable lengths, (ii) a prosodic alignment module that temporally aligns the MT output with the speech segments in the original audio, (iii) enhanced TTS to accurately control the duration of each produce utterance, and, finally, (iv) audio rendering that adds to the TTS output background noise and reverberation extracted from the original audio. In the following, we describe each component in detail, with the exception of ASR, for which we use (Di Gangi et al., 2019a) an of-the-shelf online service[2].

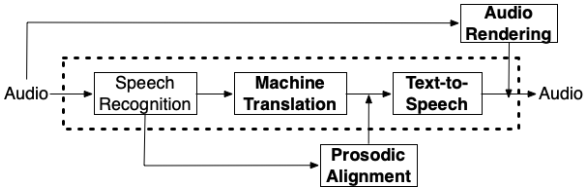

Figure 1: Speech-to-speech translation pipeline (dotted box) with enhancements to perform automatic dubbing (in bold).

3 Machine Translation

Our approach to control the length of MT output is inspired by *target forcing* in multilingual neural MT (Johnson et al., 2017; Ha et al., 2016). We partition the training sentence pairs into three groups (short, normal, long) according to the target/source string-length ratio. In practice, we select two thresholds t_1 and t_2, and partition training data according to the length-ratio intervals $[0, t_1)$, $[t_1, t_2)$ and $[t_2, \infty]$. At training time a *length token* is prepended to each source sentence according to its group, in order to let the neural MT model discriminate between the groups. At inference time, the length token is instead prepended to bias the model to generate a translation of the desired length type. We trained a Transformer model (Vaswani et al., 2017) with output length control on web crawled and proprietary data amounting to 150 million English-Italian sentence pairs (with no overlap with the test data). The model has encoder and decoder with 6 layers, layer size of 1024, hidden size of 4096 on feed forward layers, and 16

[2]Amazon Transcribe: https://aws.amazon.com/transcribe.

heads in the multi-head attention. For the reported experiments, we trained the models with thresholds $t_1 = 0.95$ and $t_2 = 1.05$ and generated at inference time translations of the shortest type, resulting, on our test set, in an average length ratio of 0.97. A reason for the length exceeding the threshold could be that for part of test data the model did not learn ways to keep the output short. A detailed account of the approach, the followed training procedure and experimental results on the same task of this paper, but using slightly different thresholds, can be found in (Lakew et al., 2019). The paper also shows that human evaluation conducted on the short translations resulted in a minor loss in quality with respect to the model without output length control. Finally, as baseline MT system for our evaluation experiments we used an online service [3]

4 Prosodic Alignment

Prosodic alignment (Öktem et al., 2019) is the problem of segmenting the target sentence to optimally match the distribution of words and pauses[4]. Let $\mathbf{e} = e_1, e_2, \ldots, e_n$ be a source sentence of n words which is segmented according to k breakpoints $1 \leq i_1 < i_2 < \ldots i_k = n$, shortly denoted with $\mathbf{i}$. Given a target sentence $\mathbf{f} = f_1, f_2, \ldots, f_m$ of m words, the goal is to find within it k corresponding breakpoints $1 \leq j_1 < j_2 < \ldots j_k = m$ (shortly denoted with $\mathbf{j}$) that maximize the probability:

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

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

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

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

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

where $Q(j,t)$ denotes the log-probability of the optimal segmentation of $\mathbf{f}$ up to position j with t break points. It is easy to show that the solution of (1) corresponds to $Q(m,k)$ and that optimal segmentation can be efficiently computed via dynamic-programming. Let $\tilde{f}_t = f_{j_{t-1}+1}, \ldots, f_{j_t}$ and $\tilde{e}_t = e_{i_{t-1}+1}, \ldots, e_{i_t}$ indicate the t-th segments of $\mathbf{f}$ and $\mathbf{e}$, respectively, we define the conditional probability of the t-th break point in $\mathbf{f}$ by:

$$\Pr(j_t \mid j_{t-1}, t) \propto \exp\left(1 - \frac{|d(\tilde{e}_t) - d(\tilde{f}_t)|}{d(\tilde{e}_t)}\right) \times \Pr(\text{br} \mid j_t, \mathbf{f}) \quad (4)$$

The first term computes the relative match in duration between the corresponding t-th segments[5], while the second term measure the linguistic plausibility of a placing a break after the j_t in $\mathbf{f}$. For this, we simply compute the following ratio of normalized language model probabilities of text windows centered on the break point, by assuming or not the presence of a pause (br) in the middle:

$$\Pr(\text{br} \mid j, \mathbf{f}) = \frac{\Pr(f_j, \text{br}, f_{j+1})^{1/3}}{\Pr(f_j, \text{br}, f_{j+1})^{1/3} + \Pr(f_j, f_{j+1})^{1/2}}$$

The rational of our model is that we want to favor split points were also TTS was trained to produce pauses. TTS was in fact trained on read speech that generally introduces pauses in correspondence of punctuation marks such as period, comma, semicolon, colon, etc. Notice that our interest, at the moment, is to produce fluent TTS speech, not to closely match the speaking style of the original speaker. In our implementation, we use a larger text window (last and first two words), we replace words with parts-of speech, and estimate the language model with KenLM (Heafield, 2011) on the training portion of the MUST-C corpus tagged with parts-of-speech using an online service[6].

5 Text To Speech

Our neural TTS system consists of two modules: a Context Generation module, which generates a context sequence from the input text, and a Neural Vocoder module, which converts the context sequence into a speech waveform. The first one is an attention-based sequence-to-sequence network (Prateek et al., 2019; Latorre et al., 2019) that predicts a Mel-spectrogram given an input text. A grapheme-to-phoneme module converts the sequence of words into a sequence of phonemes

[3] Amazon Translate: https://aws.amazon.com/translate.

[4] In this work the minimum pause interval is set to $300ms$. Pauses are detected from the time stamps produce by force-aligning audio with the transcript (Ochshorn and Hawkins, 2017).

[5] We approximate the duration $d(\cdot)$ of a segment with the sum of the lengths of its words. We plan to use better approximations in the future, e.g. the number of syllables (Öktem et al., 2019).

[6] Amazon Comprehend:https://aws.amazon.com/comprehend.

plus augmented features like punctuation marks and prosody related features derived from the text (e.g. lexical stress). For the Context Generation module, we trained speaker-dependent models on two Italian voices, male and female, with 10 and 37 hours of high quality recordings, respectively. We use the Universal Neural Vocoder introduced in (Lorenzo-Trueba et al., 2019), pre-trained with 2000 utterances per each of the 74 voices from a proprietary database.

To ensure close matching of the duration of Italian TTS output with timing information extracted from the original English audio, for each utterance we re-size the generated Mel spectrogram using spline interpolation prior to running the Neural Vocoder. We empirically observed that this method produces speech of better quality than traditional time-stretching.

6 Audio Rendering

6.1 Foreground-Background Separation

The input audio can be seen as a mixture of foreground (speech) and background (everything else) and our goal is to extract the background and add it to the dubbed speech to make it sound more real and similar to the original. Notice that in the case of TED talks, background noise is mainly coming from the audience (claps and laughs) but sometime also from the speaker, e.g. when she is explaining some functioning equipment. For the foreground-background separation task, we adapted (Giri et al., 2019; Tolooshams et al., 2020) the popular U-Net (Ronneberger et al., 2015) architecture, which is described in detail in (Jansson et al., 2017) for a music-vocal separation task. It consists of a series of down-sampling blocks, followed by one bottom convolutional layer, followed by a series of up-sampling blocks with skip connections from the down-sampling to the up-sampling blocks. Because of the down-sampling blocks, the model can compute a number of high-level features on coarser time scales, which are concatenated with the local, high-resolution features computed from the same-level up-sampling block. This concatenation results into multi-scale features for prediction. The model operates on a time-frequency representation (spectrograms) of the audio mixture and it outputs two soft ratio masks corresponding to foreground and background, respectively, which are multiplied element-wise with the mixed spectrogram, to obtain the final estimates of the two sources. Finally, the estimated spectrograms go through an inverse short-term Fourier transform block to produce raw time domain signals. The loss function used to train the model is the sum of the L_1 losses between the target and the masked input spectrograms, for the foreground and the background (Jansson et al., 2017), respectively. The model is trained with the Adam optimizer on mixed audio provided with foreground and background ground truths. Training data was created from 360 hours of clean speech from Librispeech (foreground) and 120 hours of recording taken from audioset (Gemmeke et al., 2017) (background), from which speech was filtered out using a Voice Activity Detector (VAD). Foreground and background are mixed for different signal-to-noise ratio (SNR), to generate the audio mixtures.

6.2 Re-reverberation

In this step, we estimate the environment reverberation from the original audio and apply it to the dubbed audio. Unfortunately, estimating the room impulse response (RIR) from a reverberated signal requires solving an ill-posed blind deconvolution problem. Hence, instead of estimating the RIR, we do a blind estimation of the reverberation time (RT), which is commonly used to assess the amount of room reverberation or its effects. The RT is defined as the time interval in which the energy of a steady-state sound field decays 60 dB below its initial level after switching off the excitation source. In this work we use a Maximum Likelihood Estimation (MLE) based RT estimate (see details of the method in (Löllmann et al., 2010)). Estimated RT is then used to generate a synthetic RIR using a publicly available RIR generator (Habets, 2006). This synthetic RIR is finally applied to the dubbed audio.

7 Experimental Evaluation

We evaluated our automatic dubbing architecture (Figure 1), by running perceptual evaluations in which users are asked to grade the naturalness of video clips dubbed with three configurations (see Table 1): (A) speech-to-speech translation baseline, (B) the baseline with enhanced MT and prosodic alignment, (C) the former system enhanced with audio rendering.[7] Our evaluation fo-

[7]Notice that after preliminary experiments, we decided to not evaluate the configuration *A with Prosodic Alignment*,

System	Condition
R	Original recording (reference)
A	Speech-to-speech translation (baseline)
B	A with Enhanced MT and Pros. Align.
C	B with Audio Rendering

Table 1: Evaluated dubbing conditions.

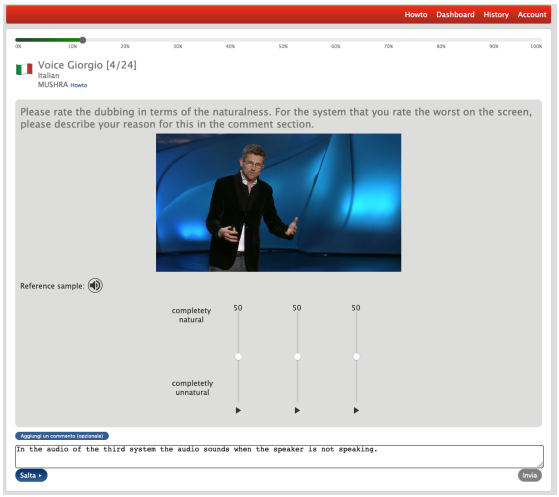

Figure 2: MUSHRA perceptual evaluation interface

cuses on two questions:

- What is the overall naturalness of automatic dubbing?
- How does each introduced enhancement contribute to the naturalness of automatic dubbing?

We adopt the MUSHRA (MUlti Stimulus test with Hidden Reference and Anchor) methodology (MUSHRA, 2014), originally designed to evaluate audio codecs and later also TTS. We asked listeners to evaluate the naturalness of each versions of a video clip on a 0-100 scale. Figure 2 shows the user interface. In absence of a human dubbed version of each clip, we decided to use, for calibration purposes, the clip in the original language as hidden reference. The clip versions to evaluate are not labeled and randomly ordered. The observer has to play each version at least once before moving forward and can leave a comment about the worse version.

In order to limit randomness introduced by ASR and TTS across the clips and by MT across versions of the same clip, we decided to run the experiments using manual speech transcripts,[8] one TTS voice per gender, and MT output by the baseline (A) and enhanced MT system (B-C) of quality judged at least acceptable by an expert.[9] With these criteria in mind, we selected 24 video clips from 6 TED Talks (3 female and 3 male speakers, 5 clips per talk) from the official test set of the MUST-C corpus (Di Gangi et al., 2019b) with the following criteria: duration of around 10-15 seconds, only one speaker talking, at least two sentences, speaker face mostly visible.

We involved in the experiment both Italian and non Italian listeners. We recommended all participants to disregard the content and only focus on the naturalness of the output. Our goal is to measure both language independent and language dependent naturalness, i.e. to verify how speech in the video resembles human speech with respect to acoustics and synchronization, and how intelligible it is to native listeners.

7.1 Results

We collected a total of 657 ratings by 14 volunteers, 5 Italian and 9 non-Italian listeners, spread over the 24 clips and three testing conditions. We conducted a statistical analysis of the data with linear mixed-effects models using the `lme4` package for R (Bates et al., 2015). We analyzed the naturalness score (response variable) against the following two-level fixed effects: dubbing system A vs. B, system A vs. C, and system B vs. C. We run separate analysis for Italian and non-Italian listeners. In our mixed models, listeners and video clips are random effects, as they represent a tiny sample of the respective true populations(Bates et al., 2015). We keep models maximal, i.e. with intercepts and slopes for each random effect, end remove terms required to avoid singularities. Each model is fitted by maximum likelihood and significance of intercepts and slopes are computed via t-test.

Table 2 summarized our results. In the first comparison, baseline (A) versus the system with enhanced MT and prosody alignment (B), we see that both non-Italian and Italian listeners perceive a similar naturalness of system A (46.81 vs.

given its very poor quality, as also reported in (Öktem et al., 2019). Other intermediate configurations were not explored to limit the workload of the subjects participating in the experiment.

[8] We would clearly expect significant drop in dubbing quality due to the propagation of ASR errors.

[9] We use the scale: 1 - Not acceptable: not fluent or not correct; 2 - Acceptable: almost fluent and almost correct; 3 - Good: fluent and correct.

47.22). When moving to system B, non-Italian listeners perceive a small improvement (+1.14), although not statistically significant, while Italian speaker perceive a statistically significant degradation (-10.93).

In the comparison between B and C (i.e. B enhanced with audio rendering), we see that non-Italian listeners observe a significant increase in naturalness (+10.34), statistically significant, while Italian listeners perceive a smaller and not statistical significant improvement (+1.05).

The final comparison between A and C gives almost consistent results with the previous two evaluations: non-Italian listeners perceive better quality in condition C (+11.01), while Italian listeners perceive lower quality (-9.60). Both variations are however not statistically significant due to the higher standard errors of the slope estimates ΔC. Notice in fact that each mixed-effects model is trained on distinct data sets and with different random effect variables. A closer look at the random effects parameters indeed shows that for the B vs. C comparison, the standard deviation estimate of the listener intercept is 3.70, while for the A vs. C one it is 11.02. In other words, much higher variability across user scores is observed in the A vs. C case rather than in the B vs. C case. A much smaller increase is instead observed across the video-clip random intercepts, i.e. from 11.80 to 12.66. The comments left by the Italian listeners tell that the main problem of system B is the unnaturalness of the speaking rate, i.e. is is either too slow, too fast, or too uneven.

The distributions of the MUSHRA scores presented at the top of Figure 3 confirm our analysis. What is more relevant, the distribution of the rank order (bottom) strengths our previous analysis. Italian listeners tend to rank system A the best system (median 1.0) and vary their preference between systems B and C (both with median 2.0). In contrast, non-Italian rank system A as the worse system (median 2.5), system B as the second (median 2.0), and statistically significantly prefer system C as the best system (median 1.0).

Hence, while our preliminary evaluation found that shorter MT output can potentially enable better synchronization, the combination of MT and prosodic alignment appears to be still problematic and prone to generate unnatural speech. In other words, while non-Italian listeners seem to value synchronicity achieved through prosodic alignment, Italian listeners seem to prefer trading synchronicity for more fluent speech. We think that more work is needed to get MT closer to the script adaptation (Chaume, 2004) style used for dubbing, and to improve the accuracy of prosodic alignment.

	Non Italian		Italian	
Fixed effects	Estim	SE	Estim.	SE
A intercept	46.81$^{\bullet}$	4.03	47.22$^{\bullet}$	6.81
ΔB slope	+1.14	4.02	-10.93*	4.70
B intercept	47.74$^{\bullet}$	3.21	35.19$^{\bullet}$	7.22
ΔC slope	+10.34^{+}	3.53	+1.05	2.30
A intercept	46.92$^{\bullet}$	4.95	45.29$^{\bullet}$	7.42
ΔC slope	+11.01	6.51	-9.60	4.89

Table 2: Summary of the analysis of the evaluation with mixed-effects models. From top down: A vs. B, B vs. C, A vs. C. For each fixed effect, we report the estimate and standard error. Symbols $^{\bullet}$, *, $^{+}$ indicate significance levels of 0.001, 0.01, and 0.05, respectively.

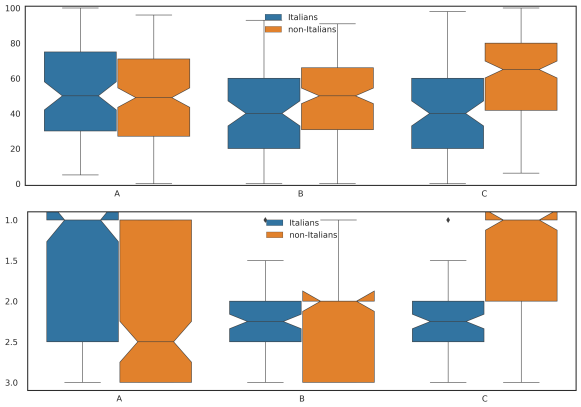

Figure 3: Boxplots with the MUSHRA scores (top) and Rank Order (bottom) per system and mother language (Italian vs Non-Italian).

The incorporation of audio rendering (system C) significantly improves the experience of the non-Italian listeners (66 in median) respect to systems B and C. This points out the relevance of including para-linguist aspects (i.e. applause's, audience laughs in jokes,etc.) and acoustic conditions (i.e. reverberation, ambient noise, etc.). For the target (Italian) listeners this improvement appears instead masked by the disfluencies introduced by the prosodic alignment step. If we try to directly measure the relative gains given by audio rendering, we see that Italian listeners score system B better than system A 27% of the times and system C better than A 31% of the times, which is a 15% relative gain. On the contrary non-Italian

speakers score B better than A 52% of the times, and C better than A 66% of the times, which is a 27% relative gain.

8 Conclusions

We have perceptually evaluated the naturalness of automatic speech dubbing after enhancing a baseline speech-to-speech translation system with the possibility to control the verbosity of the translation output, to segment and synchronize the target words with the speech-pause structure of the source utterances, and to enrich TTS speech with ambient noise and reverberation extracted from the original audio. We tested our system with both Italian and non-Italian listeners in order to evaluate both language independent and language dependent naturalness of dubbed videos. Results show that while we succeeded at achieving synchronization at the phrasal level, our prosodic alignment step negatively impacts on the fluency and prosody of the generated language. The impact of these disfluencies on native listeners seems to partially mask the effect of the audio rendering with background noise and reverberation, which instead results in a major increase of naturalness for non-Italian listeners. Future work will be devoted to better adapt machine translation to the style used in dubbing and to improve the quality of prosodic alignment, by generating more accurate sentence segmentation and by introducing more flexible synchronization.

9 Acknowledgements

The authors would like to thank the Amazon Polly, Translate and Transcribe research teams; Adam Michalski, Alessandra Brusadin, Mattia Di Gangi and Surafel Melaku for contributions to the project, and all colleagues at Amazon AWS who helped with the evaluation.

References

Douglas Bates, Martin Mchler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48.

F. Casacuberta, M. Federico, H. Ney, and E. Vidal. 2008. Recent efforts in spoken language translation. *IEEE Signal Processing Magazine*, 25(3):80–88.

Frederic Chaume. 2004. Synchronization in dubbing: A translation approach. In *Topics in Audiovisual Translation*, pages 35–52. John Benjamins B.V.

Martine Danan. 1991. Dubbing as an Expression of Nationalism. *Meta: Translators' Journal*, 36(4):606–614.

Mattia Di Gangi, Robert Enyedi, Alessndra Brusadin, and Marcello Federico. 2019a. Robust neural machine translation for clean and noisy speech translation. In *Proc. IWSLT*.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019b. MuST-C: a Multilingual Speech Translation Corpus. In *Proc. NAACL*, pages 2012–2017.

Shoichi Furukawa, Takuya Kato, Pavel Savkin, and Shigeo Morishima. 2016. Video reshuffling: automatic video dubbing without prior knowledge. In *Proc. ACM SIGGRAPH*.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. ICASSP*, pages 776–780.

Ritwik Giri, Umut Isik, and Arvindh Krishnaswamy. 2019. Attention Wave-U-Net for Speech Enhancement. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 249–253, New Paltz, NY, USA. IEEE.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *Proc. IWSLT*.

Emanuel AP Habets. 2006. Room impulse response generator. Technical Report 2.4, Technische Universiteit Eindhoven.

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

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197.

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

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Trans. of the ACL*, 5:339–351.

Richard Kilborn. 1993. 'Speak my language': current attitudes to television subtitling and dubbing. *Media, Culture & Society*, 15(4):641–660.

Cees M. Koolstra, Allerd L. Peeters, and Herman Spinhof. 2002. The Pros and Cons of Dubbing and Subtitling. *European Journal of Communication*, 17(3):325–354.

Surafel Melaku Lakew, Mattia Di Gangi, and Marcello Federico. 2019. Controlling the output length of neural machine translation. In *Proc. IWSLT*.

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

Heiner Löllmann, Emre Yilmaz, Marco Jeub, and Peter Vary. 2010. An improved algorithm for blind reverberation time estimation. In *Proc. IWAENC*, pages 1–4.

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

Xènia Martínez. 2004. Film dubbing, its process and translation. In *Topics in Audiovisual Translation*, pages 3–8. John Benjamins B.V.

J. Matoušek, Z. Hanzlìček, D. Tihelka, and M. Mèner. 2010. Automatic dubbing of TV programmes for the hearing impaired. In *Proc. IEEE Signal Processing*, pages 589–592.

J. Matoušek and J. Vít. 2012. Improving automatic dubbing with subtitle timing optimisation using video cut detection. In *Proc. ICASSP*, pages 2385–2388.

F. Metze, J. McDonough, H. Soltau, A. Waibel, A. Lavie, S. Burger, C. Langley, K. Laskowski, L. Levin, T. Schultz, F. Pianesi, R. Cattoni, G. Lazzari, N. Mana, and E. Pianta. 2002. The NESPOLE! Speech-to-speech Translation System. In *Proc. HLT*, pages 378–383.

MUSHRA. 2014. *Method for the subjective assessment of intermediate quality level of coding systems*. International Communication Union. Recommendation ITU-R BS.1534-2.

S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto. 2006. The ATR Multilingual Speech-to-Speech Translation System. *IEEE Trans. on Audio, Speech, and Language Processing*, 14(2):365–376.

R. M. Ochshorn and M. Hawkins. 2017. Gentle Forced Aligner. Https://lowerquality.com/gentle/.

Alp Öktem, Mireia Farrùs, and Antonio Bonafonte. 2019. Prosodic Phrase Alignment for Machine Dubbing. In *Proc. Interspeech*.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Proc. ICMAI*, pages 234–241. Springer.

Bahareh Tolooshams, Ritwik Giri, Andrew H. Song, Umut Isik, and Arvindh Krishnaswamy. 2020. Channel-Attention Dense U-Net for Multichannel Speech Enhancement. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 836–840, Barcelona, Spain.

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

Werner Verhelst. 1997. Automatic Post-Synchronization of Speech Utterances. In *Proc. Eurospeech*, pages 899–902.

E. Vidal. 1997. Finite-state speech-to-speech translation. In *Proc. ICASSP*, volume 1, pages 111–114 vol.1.

Wolfgang Wahlster. 2013. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer Science & Business Media. Google-Books-ID: NoqrCAAAQBAJ.

A. Waibel. 1996. Interactive translation of conversational speech. *Computer*, 29(7):41–48.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-Sequence Models Can Directly Translate Foreign Speech. In *Proc. Interspeech 2017*, pages 2625–2629. ISCA.