Robust Translation of French Live Speech Transcripts

Elise Bertin-Lemée
guillaume.klein@systrangroup.com

Guillaume Klein
josep.crego@systrangroup.com

Josep Crego
jean.senellart@systrangroup.com

Jean Senellart
SYSTRAN, 5 rue Feydeau, 75002 Paris, France

Abstract

Despite a narrowed performance gap with direct approaches, cascade solutions, involving automatic speech recognition (ASR) and machine translation (MT) are still largely employed in speech translation (ST). Direct approaches employing a single model to translate the input speech signal suffer from the critical bottleneck of data scarcity. In addition, multiple industry applications display speech transcripts alongside translations, making cascade approaches more realistic and practical. In the context of cascaded simultaneous ST, we propose several solutions to adapt a neural MT network to take as input the transcripts output by an ASR system. Adaptation is achieved by enriching speech transcripts and MT data sets so that they more closely resemble each other, thereby improving the system robustness to error propagation and enhancing result legibility for humans. We address aspects such as sentence boundaries, capitalisation, punctuation, hesitations, repetitions, homophones, etc. while taking into account the low latency requirement of simultaneous ST systems.

1 Introduction

Speech translation is the task of converting speech utterances given in a source language into text written in a different, target language. Conventional ST systems employ a two-step cascaded pipeline composed of ASR and MT modules Casacuberta et al. (2004); Waibel and Fugen (2008). One of the main drawbacks of these systems is error propagation, a problem that has received considerable attention in the last years Ruiz and Federico (2014); Sperber et al. (2017b). Multiple research efforts have tried to tightly integrate both modules by using N-best lists or word lattices Matusov et al. (2006); Dyer et al. (2008); Sperber et al. (2017a). These systems are nowadays strongly challenged by direct approaches employing a single model to translate the input speech signal, where all network components are jointly trained to maximize translation performance without the need for an intermediate readable representation Berard et al. (2016); Bansal et al. (2017); Weiss et al. (2017). Despite their architectural simplicity, reduced information loss and minimal error propagation of direct systems, cascaded solutions are still not widely used, mainly because of the data scarcity problem. Moreover, industry applications usually display speech transcripts alongside translations, making cascade approaches more realistic and practical.

Within the standard cascaded framework, researchers have encountered many challenges, mainly based on the fact that ASR transcripts exhibit very different features from those of the texts used to train neural machine translation (NMT) networks. While NMT models are often
trained with clean and well-structured text, spoken utterances contain multiple disfluencies and recognition errors which are not well modeled by NMT systems. In addition, ASR systems do not usually predict sentence boundaries or capital letters correctly, as they are not reliably accessible as acoustic cues Makhija et al. (2019); Nguyen et al. (2019). While ASR output is sufficient for many applications, where speech segments are usually short, it is difficult to use in applications that transcribe long speech segments Li et al. (2021). Typical ASR systems segment the input speech using only acoustic information, i.e., pauses in speaking, which greatly differ from the units expected by conventional MT systems. At the other end of the spectrum, systems using longer segments may span multiple sentences. This causes important translation delays, which harms the reading experience. Limited translation delays are typically achieved via starting translation before the entire audio input is received, a practice that introduces important challenges Matusov et al. (2007); Niehues et al. (2016); Arivazhagan et al. (2020).

In this work, we consider live speech-to-text translation, a task closely resembling simultaneous interpreting, that performs multilingual translations in real time and that has recently been in increasing demand in a variety of settings (radio and television broadcasts, movies, podcasts, online meetings, conferences and lectures, live events, etc.). We propose a simple but efficient ST system following a cascaded ASR-MT pipeline for live translation of French speeches into English with focus on the political discourse domain. Figure 1 shows a screenshot of our live ST system interface. Inspired by Martucci et al. (2021); Ruiz et al. (2015), we propose several data augmentation techniques to simulate errors generated by an ASR system, thus allowing the MT system to recover from ASR errors.

Figure 1: Speech translation system in action. French transcriptions and English translations are shown in real-time as they are decoded from the ASR transcripts.

Our contributions are summarised as follows:

- We detail our framework for multilingual live speech translation in the discourse domain.
- We identify discrepancies between written texts, commonly used in MT training data sets, and ASR outputs.
• To strengthen MT robustness, we propose several data augmentation methods to corrupt clean texts so as to emulate ill-formed transcripts. Notice that our approach is ASR-independent, noise introduced in the MT training can be successfully applied to errors made by other ASR systems.

• We conduct an empirical evaluation of our proposed workflows for a French-English multilingual translation task.

After introducing and presenting related work, we outline the particularities of the used speech transcripts in section 2. Details of the presented framework for live multilingual ST are given in section 3. Section 4 describes our experimental framework. Results are presented in section 5. Finally, section 6 concludes this work.

2 Speech Transcripts

A vast amount of audio sources are nowadays being produced on a daily basis. ASR systems enable such speech content to be used in multiple applications (i.e. indexing, cataloging, subtitling, translation, multimedia content production, etc). Details depend of individual ASR systems but their output, commonly called transcripts, typically consist of plain text enriched with time codes. Figure 2 (top) illustrates the transcript resulting from a French utterance. Notice time codes and confidence scores for each record. Latency records indicate pauses.

This paper focuses on French discourses, i.e. speeches delivered in reasonably good acoustic conditions and by speakers used to addressing large audiences. Under these particular conditions, we next identify the most challenging features of this kind of speeches that need to be tackled for better human or machine processing:

Sentence boundaries Speech units contained in transcripts do not always correspond to sentences as they are established in written text. Sentence boundaries provide a basis for further processing of natural language.
**Punctuation** Partially due to absence of sentence boundaries, no punctuation marks are produced by ASR systems in real time mode, a key feature for the legibility of speech transcriptions.

**Capitalisation** Transcriptions do not include correct capitalisation. A truecasing task is needed to assign each word its corresponding case information, usually depending on context.

**Number representation** Numbers provide a challenge for transcription, in particular number segmentation. See for instance the example of Figure 2 (top) where the uttered number 2001 is wrongly transcribed as a sequence of three numbers: 2, 1000 and 1. Both transcriptions may be possible, only the use of context can help to pick the right one.

**Disfluencies** Speech disfluencies such as hesitations, filled pauses, lengthened syllables, within-phrase silent pauses, repetitions are among the most frequent markers of spontaneity. Disfluencies are the most important source of discrepancies between spontaneous speech and text. Figure 2 (middle and bottom) shows transcripts with speech disfluencies.

**Recognition errors** ASR systems are error-prone. Multiple misrecognition types exist. For this work, we mainly consider errors due to homophones, missed utterances, wrongly inserted words and inflection changes. Figure 2 (middle and bottom) illustrates a transcript containing some of such errors.

### 3 Live Speech Translation

Our ST system is a standard cascading ASR-MT pipeline, where ASR outputs a French single-best hypothesis without punctuation, lower-cased, non segmented and containing multiple recognition disfluencies. To alleviate the ASR-MT mismatch we employ neural models that: (1) transform noisy ASR hypotheses into clean data (FR2fr) prior to translation (fr2en); (2) translate noisy ASR outputs (FR2en) and (3) performs both tasks at the same time, cleaning the ASR output and translation (FR2fr:en). Figure 3 illustrates the three translation pipelines implemented in this work that perform translation into English of French utterances. We use FR to indicate French transcripts while fr indicate French clean sentences.

![Figure 3: Speech translation pipelines.](image)

High quality neural models can only be learned when fed with large amounts of parallel data. Since there are scarce parallel noisy/clean resources for French, we decide to generate synthetic ASR noise from clean French texts for which English translations exist, thereby making the triplets noisy French/clean French/clean English available. In the next lines we detail the generation of different types of noise injected into clean French speeches to make them similar

---

**Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas**

Orlando, USA, September 12-16, 2022. Volume 2: Users and Providers Track and Government Track

Page 458
The palace is empty.

The palace is empty.

The palace is empty.

The palace is empty.

The palace is empty.

The palace is empty.

The palace is empty.

The palace is empty.

The palace is empty.
We tackle this problem by decoding the ASR output whenever new words become available. Figure 4 illustrates the inference steps performed by our networks when decoding the French ASR transcript “le palais est vite (pause) le roi et parti il reviens demain”. Column FR indicates, in red color, words output by the ASR at each time step $t$. Columns fr, en and fr:en indicate the corresponding output of our models (respectively FR2fr, FR2en and FR2fr:en) for the input (FR) at time step $t$. Notice that input streams remove previous sentences when an end of sentence (eos) is predicted by our model followed by $N$ words. This strikes a fair balance between flexibility and stability for segmentation choices, allowing the model to reconsider its initial prediction while ensuring consistent choices to be retained. Notice also that after predicting (eos), words output by our models consist of the same words output by the ASR, This allow us to identify the prefix to use when building new inputs (underlined strings). The prefix also contains the last token predicted for the previous sentence followed by (eos) to predict the case of the initial word of each sentence.

4 Experimental setup

| pronounced transcript | FR2fr | FR2fr+fr2en |
|-----------------------|-------|------------|
| Ces alignements là pour que le système d’intelligence artificielle fonctionne, il faut le faire sur beaucoup beaucoup de données. | Ces alignements là pour que le système d’intelligence artificielle fonctionne, il faut le faire sur beaucoup beaucoup de données. | Ces alignements là pour que le système d’intelligence artificielle fonctionne, il faut le faire sur beaucoup beaucoup de données. |
| Pronounced Transcript | J’en, on voit quand même qu’il y avait des choses qui fonctionnent pas mal. | Jean, on voit quand même qu’il y avait des choses qui fonctionnent pas mal. |
| FR2fr | FR2fr+fr2en |
| And here’s the fundamental bankruptcy of man, so fundamental that all others derive from it. thank you. | And this is where the fundamental human failure has taken place. So fundamental. Thank you for all the other things. |
| FR2fr+fr2en |
| And here occurred man’s fundamental failure, so fundamental. Thank you для all the other things. |

Figure 5: Examples where FR2fr model segments and punctuates the ASR output, correcting homophones, repetition and missing words. We observe that further work could tackle multi-word homophones or quasi-homophones and written rewording of speech-specific structures.

| Transcript | fr2en |
|------------|-------|
| et c’est ici que s’est produite la faillite fondamentale de l’homme (pause) si fondamentale que toutes les autres en découlent merci | And here’s the fundamental bankruptcy of the human, so fundamental that all other failures derive from it... thank you. |
| FR2fr | FR2fr+fr2en |
| Et c’est ici que s’est produite la faillite fondamentale de l’homme, si fondamentale que toutes les autres en découlent. merci | And this is where the fundamental human failure has taken place. So fundamental. Thank you for all the other things. |
| FR2fr+fr2en |
| And this is where the fundamental human failure has taken place. So fundamental. Thank you for all the other things. |
| Reference | [...] And here occurred man’s fundamental failure, so fundamental that all other failures ensue it... thank you. |

Figure 6: Example where FR2fr+fr2en achieves the best translation by correcting homophones and meaningfully segmenting (in red incorrect segmentations incurred by other models).

4.1 Datasets

Table 1 provides some statistics on the parallel French-English corpora employed for in this work. Statistics are computed after a light tokenization (splitting off punctuation). We employ for training available corpora close to the political discourse domain consisting on: EPPS Tiedemann (2012) (proceedings of the European Parliament), TEDX Reimers and Gurevych (2020) (subtitles of TED talks), and UNPC Ziemska et al. (2016) (official records and documents of the United Nations Parliament). For testing we use the testsets from two multilingual ST corpus,
EPST Iranzo-Sánchez et al. (2020)(Europarl ST) and MTEDXSalesky et al. (2021) (Multilingual TEDx). All data is pre-processed using the OpenNMT tokenizer\(^3\).

| Corpus | Sentences | Words | Vocab |
|--------|-----------|-------|-------|
|        | En Fr     |       |       |
| Train  |           |       |       |
| EPPS   | 2.1M      | 57.7M | 66.7M |
| TEDX   | 0.4M      | 8.4M  | 9.1M  |
| UNPC   | 30.3M     | 792.5M| 1016.3M|
|        |           |       |       |
| Test   | 1804      | 50k   | 55k   |
| 18k    | 21k       | 3.1k  | 3.4k  |

Table 1: Statistics of parallel corpora used for train and test sets.

### 4.2 Network and Training Details

All our models follow the Transformer architecture Vaswani et al. (2017) implemented by the OpenNMT-tf\(^4\) toolkit. More precisely, our fr2en, FR2fr, FR2en and FR2fr:en models use: Word embedding size: 1024; Number of layers: 6; Number of heads in multi-head self-attention layer: 16; Inner dimension of feedforward layer: 4096; Dropout rate: 0.1. Our FR2fr model uses a smaller version of the same architecture with: Word embedding size: 512; Number of layers: 4; Number of heads in multi-head self-attention layer: 8; Inner dimension of feedforward layer: 1024; In all cases, we use shared embeddings for both the input and output layers. The encoder and decoder use the same BPE units learned from source and target corpora with 16,000 merge operations. Learning is performed over 1 GPU during 300K steps with a batch size of 64K tokens per step. We applied label smoothing to the cross-entropy loss with a rate of 0.1. Resulting models are built after averaging the last five checkpoints of the training process.

In order to build our FR2xx models we need parallel speeches rather than parallel sentences: to simulate consecutive sentences we join lists of 5 to 25 random sentences of the corpora. Note that inter-sentence context is only employed by our models to predict the case of the initial word of each sentence. All our experiments use ASR transcripts produced by VoxSigma web service API by Vocapia Research\(^5\), a state-of-the-art neural ASR system for French language.

### 5 Experimental Results

Table 2 indicates BLEU\(^6\) accuracy results of our three different pipelines as detailed in Figure 3 as well as the fr2en system that is trained on clean parallel texts.

As it can be seen, the fr2en model, trained on clean parallel data, exhibits the worst results. Differences in training and inference data sets significantly impact performance. Concerning models learned using noisy source data, best BLEU performance is achieved by the FR2en model. We hypothesize that FR2fr+fr2en suffers from error propagation, which means errors introduced in the first module FR2fr can not be recovered by the fr2en module. Results by FR2fr:en are very similar to those obtained by FR2fr+fr2en. Despite its

\(^3\)https://github.com/OpenNMT/Tokenizer
\(^4\)https://github.com/OpenNMT/OpenNMT-tf
\(^5\)https://www.vocapia.com/voxsigma-speech-to-text.html
\(^6\)https://github.com/mjpost/sacrebleu
lower BLEU score, the model **FR2fr+fr2en** outputs clean **fr** transcripts, which is an important asset for some industry applications, and can impact segmentation and translation, as can be seen in Figure 6.

| System               | Europarl ST | MTEDX |
|----------------------|-------------|-------|
| **fr2en**            | 28.56       | 23.20 |
| **FR2fr+fr2en**      | 32.65       | 29.53 |
| **FR2en**            | 35.21       | 32.86 |
| **FR2fr:en**         | 31.80       | 29.87 |

Table 2: BLEU score on Europarl ST and Multilingual TEDx testset

The presented framework delivers translations with very low delay rates. Each new word supplied by the ASR produces a new translation hypothesis which is immediately displayed to the user. Even though the segment being decoded can fluctuate (translation changes when including additional words), as soon as an end of segment (**eos**) followed by a fixed number of words is predicted, the segment remains unchanged. We encountered very limited fluctuations, impacting the last words of the hypotheses being decoded.

### 6 Conclusions

We presented a framework for live speech translation based on the cascaded approach. We proposed several techniques to automatically enrich clean parallel corpora with several noise types typically present in speech transcripts, thereby improving the system robustness to error propagation. We pay special attention to translation delay rates to enhance legibility for humans. Results indicate the suitability of the framework presented showing important accuracy gains when compared to a baseline system and attaining very low delay rates. We plan to extend this work using a transformer with dual decoder: a system that uses a single encoder for the ASR transcripts and two parallel decoders to produce a clean version of the transcript in the same language and its corresponding translation, with the ability to attend to each other. This way, we expect to obtain similar delay rates with improved translation accuracy than our best performing model, and to additionally produce clean transcripts.

### Acknowledgements

The work presented in this paper was partially supported by the European Commission under contract H2020-787061 ANITA.

### References

Arivazhagan, N., Cherry, C., Te, I., Macherey, W., Baljekar, P., and Foster, G. (2020). Re-translation strategies for long form, simultaneous, spoken language translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7919–7923. IEEE.

Bansal, S., Kamper, H., Lopez, A., and Goldwater, S. (2017). Towards speech-to-text translation without speech recognition. *CoRR*, abs/1702.03856.

Berard, A., Pietquin, O., Servan, C., and Besacier, L. (2016). Listen and translate: A proof of concept for end-to-end speech-to-text translation. *ArXiv*, abs/1612.01744.

Casacuberta, F., Ney, H., Och, F., Vidal, E., Vilar, J., Barrachina, S., García-Varea, I., Llorens, D., Martínez, C., Molau, S., Nevada, F., Pastor, M., Picó, D., Sanchis, A., and Tillmann, C. (2004). Some
approaches to statistical and finite-state speech-to-speech translation. *Computer Speech & Language*, 18(1):25–47.

Dyer, C., Muresan, S., and Resnik, P. (2008). Generalizing word lattice translation. In *Proceedings of ACL-08: HLT*, pages 1012–1020, Columbus, Ohio. Association for Computational Linguistics.

Iranzo-Sánchez, J., Silvestre-Cerdà, J. A., Jorge, J., Roselló, N., Giménez, A., Sanchis, A., Civera, J., and Juan, A. (2020). Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233.

Li, D., Te, I., Arivazhagan, N., Cherry, C., and Padfield, D. (2021). Sentence boundary augmentation for neural machine translation robustness. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7553–7557. IEEE.

Makhija, K., Ho, T.-N., and Chng, E.-S. (2019). Transfer learning for punctuation prediction. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 268–273.

Martucci, G., Cettolo, M., Negri, M., and Turchi, M. (2021). Lexical Modeling of ASR Errors for Robust Speech Translation. In *Proc. Interspeech 2021*, pages 2282–2286.

Matusov, E., Hillard, D., Magimai-Doss, M., Hakkani-Tur, D., Ostendorf, M., and Ney, H. (2007). Improving speech translation with automatic boundary prediction. In *8th Annual Conference of the International Speech Communication Association (INTERSPEECH 2007)*, pages 2449–2452.

Matusov, E., Kanthak, S., and Ney, H. (2006). Integrating speech recognition and machine translation: Where do we stand? In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5, pages V–V.

Nguyen, B., Nguyen, V. B. H., Nguyen, H., Phuong, P. N., Nguyen, T., Do, Q. T., and Mai, L. C. (2019). Fast and accurate capitalization and punctuation for automatic speech recognition using transformer and chunk merging. *CoRR*, abs/1908.02404.

Niehues, J., Nguyen, T. S., Cho, E., Ha, T.-L., Kilgour, K., Müller, M., Sperber, M., Stüker, S., and Waibel, A. H. (2016). Dynamic transcription for low-latency speech translation. In *INTERSPEECH*.

Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics.

Ruiz, N. and Federico, M. (2014). Assessing the impact of speech recognition errors on machine translation quality. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track*, pages 261–274, Vancouver, Canada. Association for Machine Translation in the Americas.

Ruiz, N., Gao, Q., Lewis, W., and Federico, M. (2015). Adapting machine translation models toward misrecognized speech with text-to-speech pronunciation rules and acoustic confusability. In *Proceedings of Interspeech 2015*.

Salesky, E., Wiesner, M., Bremerman, J., Cattoni, R., Negri, M., Turchi, M., Oard, D. W., and Post, M. (2021). Multilingual tedx corpus for speech recognition and translation.
Sperber, M., Neubig, G., Niehues, J., and Waibel, A. (2017a). Neural lattice-to-sequence models for uncertain inputs. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1380–1389, Copenhagen, Denmark. Association for Computational Linguistics.

Sperber, M., Niehues, J., and Waibel, A. H. (2017b). Toward robust neural machine translation for noisy input sequences. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Chair), N. C. C., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12), Istanbul, Turkey. European Language Resources Association (ELRA).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, Advances in Neural Information Processing Systems 30, pages 5998–6008. Curran Associates, Inc.

Waibel, A. and Fugen, C. (2008). Spoken language translation. IEEE Signal Processing Magazine, 25(3):70–79.

Weiss, R. J., Chorowski, J., Jaitly, N., Wu, Y., and Chen, Z. (2017). Sequence-to-sequence models can directly translate foreign speech.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The United Nations parallel corpus v1.0. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).