Role of Pausing in Text-to-Speech Synthesis for Simultaneous Interpretation

Vivek Kumar Rangarajan Sridhar, John Chen, Srinivas Bangalore, Alistair Conkie

AT&T Labs - Research
180 Park Avenue, Florham Park, NJ 07932, USA

Abstract

The goal of simultaneous speech-to-speech (S2S) translation is to translate source language speech into target language with low latency. While conventional speech-to-speech (S2S) translation systems typically ignore the source language acoustic-prosodic information such as pausing, exploiting such information for simultaneous S2S translation can potentially aid in the chunking of source text into short phrases that can be subsequently translated incrementally with low latency. Such an approach is often used by human interpreters in simultaneous interpretation. In this work we investigate the phenomena of pausing in simultaneous interpretation and study the impact of utilizing such information for target language text-to-speech synthesis in a simultaneous S2S system. On one hand, we superimpose the source language pause information obtained through forced alignment (or decoding) in an isomorphic manner on the target side while on the other hand, we use a classifier to predict the pause information for the target text by exploiting features from the target language, source language or both. We contrast our approach with the baseline that does not use any pauses. We perform our investigation on a simultaneous interpretation corpus of Parliamentary speeches and present subjective evaluation results based on the quality of synthesized target speech.

Index Terms: simultaneous interpretation, translation, pausing, prosody, mean opinion score (MOS)

1. Introduction

Simultaneous interpretation (SI) refers to the challenging task of listening to speech in the source language and simultaneously interpreting (non-verbatim translation) it in the target language. Even though simultaneous interpreters have been providing satisfactory services daily in dozens of languages and thousands of meetings across the world (e.g., United Nations, embassies, etc.), it is an arcane art that has received little attention from the speech and language research community. One of the critical constraints in SI is that the delay between a source language chunk and its corresponding target language chunk (referred to as *ear-voice-span*) is kept minimal in order to continually engage the listeners. Simultaneous interpreters are able to generate target speech incrementally with very low ear-voice span by using a variety of strategies [1] such as anticipation, cognitive and linguistic inference, paraphrasing, etc. As a consequence, the translated segments can range from short phrases to a complete sentence.

Simultaneous translation using speech translation technology has been gradually trying to reduce the dependence on human interpreters to improve the scalability as well as eliminate the fatigue associated with prolonged human interpretation. However, target language synthesis in such systems is either ignored; i.e., only speech-to-text is enabled, or performed at the sentence level using the translated text. The notion of an utterance is typically obtained by predicting punctuation on the source text, translating the sentence and subsequently synthesizing the complete sentence using text-to-speech synthesis. Such an approach loses the rich information contained in the source speech signal that may be vital for incremental translation. Simultaneous Interpreters use several acoustic and prosodic cues from the source speech to perform linguistic inference as well as control the pace of speech production in the target language [1]; e.g., taking a breath or perform planning during a source language pause, pausing in the target language to wait for the verb in the source language, etc. Disregarding such information, especially in speech-to-speech (S2S) translation of long speeches (talks and lectures), may result in monotonous speech synthesis of long segments that may impair the understanding of target speech.

In this work we investigate the phenomena of pausing in simultaneous interpretation and examine the impact of utilizing such information for target language text-to-speech synthesis in a simultaneous S2S system. We contrast different strategies for incorporating pause information in the target language. On one hand, we superimpose the source language pause information obtained through forced alignment (or decoding) in an isomorphic manner on the target side while on the other hand, we use a classifier to predict the pause information for the target text by exploiting features from the target language, source language or both. We perform our investigation on a simultaneous interpretation corpus of Parliamentary speeches and present subjective evaluation results based on the quality of synthesized target speech.

The rest of the paper is organized as follows. In Section 2 we formally define the problem and describe the data used in this work in Section 3. We describe the experimental setup in Section 4 followed by results of the experiments in Section 4.3. We provide a brief discussion about the experimental results in Section 5 followed by conclusions and directions for future work in Section 6.

2. Problem Formulation

The basic problem of text translation can be formulated as follows. Given a source (French) sentence $\mathbf{f} = f_1^J = f_1, \cdots, f_J$, we aim to translate it into target (English) sentence $\hat{\mathbf{e}} = \hat{e}_1^I = \hat{e}_1, \cdots, \hat{e}_I$.

$$\hat{\mathbf{e}}(\mathbf{f}) = \arg\max_{\mathbf{e}} \Pr(\mathbf{e}|\mathbf{f}) \tag{1}$$

If, as in talks, the source text (reference or ASR hypothesis) is very long, i.e., J is large, we attempt to break down the source string into shorter sequences, $\mathbf{S} = s_1 \cdots s_k \cdots s_{Q_s}$, where each sequence $s_k = [f_{j_k} f_{j_k+1} \cdots f_{j_{(k+1)}-1}]$, $j_1 = 1, j_{Q_s+1} = J+1$. Let the translation (or interpretation) of each foreign sequence s_k be denoted by $t_k = [e_{i_k} e_{i_k+1} \cdots e_{i_{(k+1)}-1}]$, $i_1 =$

$1, i_{Q_s+1} = I^{'} + 1$[1]. The segmented sequences can be translated using a variety of techniques [2] while the segmentation itself can be obtained using linguistic and non-linguistic strategies [3, 4, 5]. The translated sequence, $\mathbf{T} = t_1 \cdots t_k \cdots t_{Q_s}$, is typically synthesized independently using a text-to-speech synthesizer that generates appropriate prosody and pausing using pre-trained models.

Our objective is to improve the quality of speech synthesis in the above framework by predicting pausing information for the translated sequence $\mathbf{T}$; i.e., for the output sequence $t_1 \cdots t_{Q_s} = [e_1 \cdots e_{I'+1}]$, we predict the presence or absence of silence (binned into N intervals) between each pair of words. Subsequently, the new silence inserted sequence $[e_1\ sil_1\ e_2\ sil_2\ e_3\ nosil_3 \cdots sil_{I'}\ e_{I'+1}]$ is used by the TTS engine; $sil_1, sil_2, nosil_3, sil_{I'}$ are the predicted classes in the example. Since we can get the word alignment information of a partially translated sequence, it is feasible to bootstrap source language silence information (obtained from a speech recognizer) as well as other possible syntactic information associated at a word level in the target language prediction. In training a classifier to predict pauses for the target language, one can use a variety of target as well as source language features, thus, facilitating inference from the source language signal.

We use a maximum entropy classifier for predicting the silence class after each target word. Given a sequence of translated words $e_1 \cdots e_{I'+1}$, their parts of speech (POS) $p_1 \cdots p_{I'+1}$, their corresponding source words $f_1 \cdots f_{J+1}$, and a pause label vocabulary ($l_i \,\epsilon\, \mathcal{L}, |\mathcal{L}| = N + 1$), the best pause label sequence $L^* = l_1, l_2, \cdots, l_{I'}$ is obtained by approximating the sequence classification problem, using conditional independence assumptions, to a product of local classification problems as shown in Eq.(3). The classifier is then used to assign a pause label to each target word conditioned on a vector of local contextual features from both source and target sides.

$$L^* = \arg\max_L P(L|e_1 \cdots e_{I'+1}, p_1 \cdots p_{I'+1}, f_1 \cdots f_{J+1}) \tag{2}$$

$$\approx \arg\max_L \prod_{i=1}^{n} p(l_i|e_1 \cdots e_{I'+1}, p_1 \cdots p_{I'+1}, f_1 \cdots f_{J+1}) \tag{3}$$

$$= \arg\max_L \prod_{i=1}^{n} p\left(l_i|\mathbf{\Phi}_i(e_1 \cdots e_{I'+1}, p_1 \cdots p_{I'+1}, f_1 \cdots f_{J+1})\right) \tag{4}$$

where $\mathbf{\Phi}_i(e_1 \cdots e_{I'+1}, p_1 \cdots p_{I'+1}, f_1 \cdots f_{J+1})$ is a set of features extracted within a bounded local context around word e_i.

In order to obtain POS tags for words $e_1 \cdots e_{I'+1}$, a unigram POS tagger was implemented which used word shape features to predict the POS of unknown words. The English tagger was trained on the Penn Treebank while the Spanish tagger was trained on EPIC corpus (Section 3) tagged using Spanish Freeling [6].

3. Data

In order to train the target language pause classifier, one needs a corpus that contains source speech and its corresponding target speech (either translation or interpretation). We used the European Parliamentary interpretation corpus (EPIC). The EPIC corpus [7] is a parallel corpus of European Parliamentary speeches and their corresponding simultaneous interpretations. The source speeches are either in English (81), Spanish (21) or Italian (17) and each source speech is simultaneously interpreted in two other languages. We extracted the audio from the video clips of each source language speaker while the audio for the interpreted target speeches was already provided. The corpus also contains the transcripts of all the speeches. We use only the English-Spanish portion of the corpus; i.e., the 81 speeches interpreting from English to Spanish and 21 speeches with interpretation from Spanish to English. The genre of the speeches is also provided with the corpus and can be read, impromptu or spontaneous. We picked one speech from each of these categories for testing and used the remaining for training.

As a first step in our analysis we forced aligned the English and Spanish speeches independently using generic acoustic models. The English acoustic model was trained on about 600 hours of TED talks while the Spanish acoustic model was trained on close to 1000 hours of speech collected through smartphones. Both the acoustic models were trained using minimum phone error (MPE) criterion using the AT&T WATSON$^{\text{SM}}$ speech recognizer [8]. The resulting word segmentation contained the start and end duration for each word as well as silences (with duration). Subsequently, we aligned the transcripts in the parallel speeches at the sentence level using dynamic programming with an English-Spanish dictionary.

3.1. Inducing word alignment

Unlike parallel text used in building word and phrase-based machine translation models, SI texts maybe non-parallel and even non-comparable. As a result, inducing word correspondence using automatic word alignment is quite difficult. First, we used a sentence matching algorithm [9] to align the sentences across the two languages. Subsequently, we used a custom algorithm for aligning the words across the two languages. The matching was facilitated by a dictionary obtained through automatic alignment [10] of a large English-Spanish parallel corpus comprising of about 8 million sentence pairs. The resulting dictionary was filtered such that only top 10 target translations (sorted by posterior probability) of each source word was preserved in the final dictionary.

Our word alignment procedure links each source word with its closest matching target word, if possible, according to heuristics. These heuristics take into account the amount of time between when the source word is spoken and its corresponding target word is spoken, as well as translation probabilities obtained through the dictionary. Specifically, the input consists of a sequence of source words $(f_1, f_2, \ldots, f_J)$ and a corresponding sequence of target words $(e_1, e_2, \ldots, e_I)$. In addition, there is a function TIME that maps a source or target word to its start time and another function STOP that maps a source or target word to *true* if it is a stopword and*false* otherwise. Finally, it is assumed that translation probabilities $P(e_i|f_j)$ are available.

The procedure takes three parameters: δl and δr define the left and right part of the time window in which the target word e_i corresponding to the source word f_j is taken to appear. t is a probability threshold that forbids a target word e_i from linking to a source word f_j when $P(e_i|f_j) < t$. For our experiments, we chose $\delta l = 1$ second, $\delta r = 6$ seconds, and $t = 0.008$. The procedure tries to link each source word $f_j \in (f_1, \ldots, f_J)$ to a target word as follows. First, a candidate set F_e of target words is constructed such that $e_i \in (e_1, \ldots, e_I)$ is placed in F_f if and

[1] The segmented and unsegmented talk may not be equal in length, i.e., $I \neq I^{'}$

only if the following criteria hold:

- TIME$(f_j) - \delta l \leq$ TIME$(e_i) \leq$ TIME$(f_j) + \delta r$
- STOP$(f_j) \wedge$ STOP(e_i) or $\neg$STOP$(f_j) \wedge \neg$STOP(e_i)
- $P(e_i|f_j) \geq t$

Finally, e_i^* is output where,

$$e_i^* = \arg\max_{e_i \in F_f} P(e_i|f_j) \tag{5}$$

4. Experimental Setup

We examine the utility of predicting pauses in target language for improved text-to-speech synthesis using five different stimuli. The stimuli used in our investigation is as follows.

s1: Target text separated by reference punctuation (only period)

s2: Target text with pauses obtained through forced alignment of reference target text

s3: Target text with pauses superimposed from forced alignment of reference source text

s4: Target text with pauses predicted using a classifier trained on target language features

s5: Target text with pauses predicted using a classifier trained on source and target language features

In the first stimulus **s1**, manual transcription of the interpreted speech marked with sentence boundaries is used for synthesis. We only use periods as markers of sentence boundary. In simultaneous speech-to-speech translation systems, one typically gets such an output albeit with errors introduced during automatic speech translation. In the second stimulus **s2**, we take the forced alignment of the target text obtained by using a speech recognizer and insert pauses into the text as determined by the ASR; i.e., the pausing is identical to that used by the interpreter during the target speech production. The stimulus **s3** is an isomorphic mapping of pauses from the source to the target. We project the silences obtained through forced alignment of the source speech onto the target through the word alignment procedure described in Section 3.1. Since, the interpretation procedure does not generate a perfectly parallel text, some of the words in source and target may be unaligned. We superimpose the silences only on words that are aligned using our alignment procedure.

The stimuli **s4** and **s5** are created by inserting pauses predicted automatically through a classifier. Classifiers for both **s4** and **s5** predict pauses using the following pause label vocabulary:

Label	**Meaning**
no silence	$0 \leq$ pause < 0.2 sec
short break	0.2 sec $\leq$ pause < 0.5 sec
long break	$0.5 \leq$ pause

Table 1: Description of the classes used in the classifier

Pauses in the EPIC corpus were mapped to these pause labels as follows: pauses less than 0.2 seconds were mapped to *no silence*; pauses between 0.2 and 0.5 seconds were mapped to *short break*; and pauses greater than 0.5 seconds were mapped to *long break*.

Feature sets $\mathbf{\Phi}_i$ for classifiers for both **s4** and **s5** contain words and POS in a five word window around the target word e_i to be tagged. In addition, feature set $\mathbf{\Phi}_i$ for the classifier for **s5** contained two features encoding the types of pauses, if any, that occurred before and after source word f_i to which the target word e_i has been linked.

Classifiers for **s4** and **s5** were trained on 18 speeches (source: Spanish) from the EPIC corpus and tested on 3 other speeches of this type. Results are shown below in Table reftable:classification.

	Class	*Recall*	*Precision*	*F*
s4	*no silence*	0.9811	0.8630	0.9182
	short break	0.0374	0.2667	0.0656
	long break	0.1452	0.3600	0.2069
s5	*no silence*	0.9821	0.8631	0.9188
	short break	0.1294	0.3333	0.0960
	long break	0.1452	0.4286	0.2169

Table 2: Classification performance of the classifiers used for generating stimuli **s4** and **s5**

Overall, the classification results indicate that it is quite difficult to predict short and long breaks in comparison with absence of silence. Classifier **s5** performs somewhat better than **s4**, showing that silence information from the source speech helps predict silence in the target. **s5** encoded only a small amount of such information as features; adding more information from the source speech may improve the classifier's accuracy further. In addition, the results may be skewed because the training data for our classifier is quite sparse. There were only 18 speeches interpreted from Spanish-English. As part of our current study, we are performing experiments for English-Spanish that has larger amounts of training data but require Spanish speakers to take the listening tests.

4.1. Experimental Design

The Web-based listening tests were administered in two ways: Web interface hosted on a standalone server and Amazon Mechanical Turk. We picked three speeches from the EPIC corpus; Spanish source speech interpreted into English as we had access to more English speakers for subjective listening tests. The three speeches belonged to read, impromptu and mixed genre categories to cover varying styles of the speeches. Since the source speeches were 1.5 minutes long, it was deemed that using the entire speech was too cumbersome for a listener to listen to during a listening test. Hence, we selected two 30 second snippets from each speech. The final listening test was comprised of 6 audio snippets across the five stimuli.

The listening test had 6 sections with each section comprising 5 stimuli. The listeners were asked to rate each audio file on a scale of 1-5 (bad, poor, fair, good, excellent). The listeners also indicated whether or not English was their native language, and whether they listened using headphones or speakers.

4.2. Listeners

A total of 100 listeners participated in the subjective listening test; 74 were native English speakers while 26 were non-native English speakers. Furthermore, 88 listeners took the test using headphones and the remaining 12 used their PC speakers. The average time taken for the test was 19 minutes (the minimum time to listen to all the stimuli is 30*0.5minutes=15 minutes).

4.3. Experimental Results

The results of the subjective listening test is summarized in Table 3. The table shows the mean and standard deviation of the ratings overall as well as across the 3 genres of speech (read, mixed and impromptu). The results indicate that the listeners prefer the synthesized audio from reference punctuation for the target text. However, the average length of a sentence in the test set is 19 words which is prohibitively long for synthesis in simultaneous S2S interpretation or translation. The average length of a sentence for **s2**, **s3**, **s4**, **s5** is 3, 4, 8 and 7 words, respectively. The quality of synthesis for long sentences is presumably better as the TTS engine can use longer units as well better prosody. The quality of synthesis for the other stimuli is mostly fair but significantly poorer than stimulus **s1**. It is also interesting that the quality for impromptu speech is better than that for the read and mixed mode of speech. When the speech is unplanned and more informal, the pauses predicted by the classifier are acceptable to the listener in contrast with read speech that has typically has a rigs syntactic structure. The results in general indicate that pauses either superimposed from the source speech or predicted using a classifier (target or source and target features) can offer a reasonable means of synthesizing target speech incrementally in a S2S translation setting. Considering that stimulus **s1** cannot be used in a real-time translation scenario, we need to balance latency versus synthesis quality using the approaches presented through stimuli **s2-s5**.

Stimulus	Rating (mean and standard deviation)			
	Overall	Read	Mixed	Impromptu
s1	3.6±0.9	3.5±0.9	3.7±0.8	3.6±0.9
s2	2.9±1.0	2.7±1.1	2.8±1.0	3.1±0.9
s3	2.9±1.0	2.7±1.0	3.0±1.0	3.1±0.9
s4	2.8±1.0	2.8±1.0	2.7±1.0	3.0±0.9
s5	2.9±1.0	2.9±1.0	2.9±0.9	3.0±1.0

Table 3: Mean and standard deviation of ratings across the five stimuli

5. Discussion

The experiments performed in this work are on reference text; i.e., no translation system was used for translating the source text into target language. Hence, it is the ideal case scenario where one can assume perfect translation (or interpretation). The accuracy of the pause classifier is bound to degrade while operating with noisy text translations. We plan to perform this investigation as part of future work.

The pause classifier predicts non-pauses reasonably well, but predicts pauses with poor accuracy. Part of the reason is the small amount of training data (18 speeches, about 41,500 words). Also, within this data there are about 9,500 examples of non-pauses and only 1,500 examples of pauses, which may explain the impoverished accuracy of pause prediction. Boosting methods which may better delineate the decision boundary for pauses may bring up their accuracy. In the case of **s4** since the prediction is based only on target text, one can conceivably use a large amount of non interpretation data to learn the model. However, the model is likely to predict pauses as in prepared speeches in contrast with simultaneously interpreted target speech.

Another problem with the prediction of pauses is that instead of having several local maxima in the distribution of sorted pauses in the training data to which one might assign discrete pause labels such as *short pause* or *long pause*, the distribution is a smooth curve that exponentially decreases as pause time increases. Thus, the binning of pauses into discrete labels that were done for these experiments were somewhat arbitrary.

The training data used to train the pause classifier is limited in this work as we had only 18 speeches from Spanish-English. We are currently performing experiments for Ensligh-Spanish with larger amount of training data (81 speeches). It can be expected that the classifier accuracy will increase with larger amounts of training data.

6. Conclusion

In this work we investigated the phenomena of pausing in simultaneous speech interpretation and studied the problem of using such information for target language text-to-speech synthesis in a simultaneous speech-to-speech translation system. We contrasted several ways of predicting pauses in the target language for a speech-to-speech translation setting, particularly, speech interpretation from Spanish-English. Our results indicate that either superimposing source language pauses or predicting pauses for the target language by exploiting lexical and syntactic features (both source and target language) can result in reasonably good quality synthesized speech when the input speech is unplanned, i.e., impromptu. However, the quality of synthesis suffers when the input speech is read as the speaker pauses less often. Our results also indicate that pauses can be used as good markers for chunking the source speech to reduce the latency in speech-to-speech translation. We are currently performing experiments on a larger corpus as well as analysis in English-Spanish (resulting in Spanish text-to-speech synthesis).

7. References

[1] G. V. Chernov, *Inference and anticipation in simultaneous interpreting*. John Benjamins, 2004.

[2] S. Bangalore, V. K. Rangarajan Sridhar, P. Kolan, L. Golipour, and A. Jimenez, “Real-time incremental speech-to-speech translation of dialogs,” in *Proceedings of NAACL:HLT*, June 2012.

[3] M. Cettolo and M. Federico, “Text segmentation criteria for statistical machine translation,” in *Proceedings of the 5th international conference on Advances in Natural Language Processing*, 2006.

[4] C. Fügen, A. Waibel, and M. Kolss, “Simultaneous translation of lectures and speeches,” *Machine Translation*, vol. 21, pp. 209–252, 2007.

[5] C. Fügen and M. Kolss, “The influence of utterance chunking on machine translation performance,” in *Proceedings of Interspeech*, 2007.

[6] L. Padr and E. Stanilovsky, “Freeling 3.0: Towards wider multilinguality,” in *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*. Istanbul, Turkey: ELRA, May 2012.

[7] C. Bendazzoli and A. Sandrelli, “An approach to corpus-based interpreting studies,” in *Proceedings of the Marie Curie Euroconferences MuTra: Challenges of Multidimensional Translation*, Saarbrücken, 2005.

[8] V. Goffin, C. Allauzen, E. Bocchieri, D. Hakkani-Tür, A. Ljolje, and S. Parthasarathy, “The AT&T Watson Speech Recognizer,” Tech. Rep., September 2004.

[9] V. K. Rangarajan Sridhar, L. Barbosa, and S. Bangalore, “A scalable approach to building a parallel corpus from the Web,” in *Proceedings of Interspeech*, 2011.

[10] F. J. Och and H. Ney, “A systematic comparison of various statistical alignment models,” *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.