Pause Transfer in the Speech-to-Speech Translation Domain

Pablo Daniel Agüero and Juan Carlos Tulli

Communications Lab
University of Mar del Plata
Argentina

Antonio Bonafonte

TALP Research Center
Universitat Politècnica de Catalunya
Spain

Abstract

In the speech-to-speech translation framework automatic speech recognition and spoken language translation components provide additional information about the location of pauses in the source language. This information may be useful to improve the performance of pause prediction algorithms for speech synthesis. In this paper we propose a transfer algorithm based on tuples. The results show a better performance of the proposed approach with respect to a baseline system that does not use source language information.

1. Introduction

Phrasing is a very important topic to achieve high-quality speech synthesis. It consists on breaking long sentences into smaller prosodic phrases using several acoustics cues: pauses, tonal changes, lengthening of the final syllable, etc.

The prosodic boundaries have several uses: semantic, pragmatic and physiological.

One of the semantic usages is related to the disambiguation of the meaning of a sentence or a part of it. A misplaced phrase break may radically change the meaning of a sentence. For example, in the sentence

"The plot concerns the guardian of the prince who was exiled from the country for decades."

the meaning is different according to the position of the phrase break:

"The plot concerns the guardian <phrase break> of the prince who was exiled from the country for decades." ***Meaning: The prince is exiled.***

*"The plot concerns the guardian of the prince <phrase break> who was exiled from the country for decades."****Meaning: The guardian of the prince is exiled.***

Concerning to pragmatics, the use of phrase breaks may introduce new information in the sentence to convey a different meaning. For example, a longer pause may give more importance to a particular portion of an utterance changing the original interpretation of the sentence without such pause.

Finally, prosodic boundaries with pauses are necessary for breathing. The distance between pauses depends on many factors, such as speech rate, physiological condition of the speaker, discourse structure, etc.

TC-STAR (Technology and Corpora for Speech to Speech Translation) was financed by European Commission within the Sixth Program. It was envisaged as a long-term effort to advance research in all core technologies for Speech-to-Speech Translation (SST): Automatic Speech Recognition (ASR), Spoken Language Translation (SLT) and Text to Speech (TTS) (speech synthesis).

One of the goals of this project was the use of acoustic parameters of the source speaker to help in the generation of the prosody of the target speaker by combining information from ASR, SLT and TTS components in the parliamentary domain, as shown in Figure 1. Word boundaries provided by ASR are used to extract acoustic information from each word. SLT component has available alignment links to relate the words in source and target language. Therefore, it is posible to use the source speech to influence the prosodic output of the speech synthesis component.

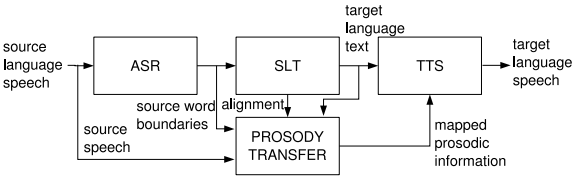


Figure 1: Prosody transfer in the framework of speech-to-speech translation

Within this project pause transfer of the source speaker is a key issue because it conveys information about the style of the discourse and important semantic information to disambiguate the meaning of the utterance and the intended meaning of the speaker. This is crucial in a parliament to avoid misinterpretations.

In this paper we will review two approaches to generate the pauses for speech synthesis in the framework of speech-to-speech translation.

In the first approach we use a state-of-the-art approach to predict pauses without using the pauses of the source speaker.

The second approach is a step forward. We apply a technique to transfer the pauses of the source speaker both in position and duration to enhance the quality of the speech synthesis, resemble the style of the source speaker and keep the meaning and intention of the discourse.

In Section 2 we will explain the baseline system used in our experiments. Section 3 has a description of the proposed approach. The experimental results are shown in Section 4. Finally, the conclusions and future work are in Section 5.

2. Baseline system

Several data-driven approaches have been proposed in the literature to predict phrase break boundaries.

Prieto et al. [1] proposed to train a decision tree to place phrase boundaries. Kohen et al. [2] made a modification to the

previous approach adding syntactic features, reporting a significant improvement.

These two methods place boundaries taking into account local information. They do not use the location of previous boundaries on the decision. Black et al. [3] have proposed a different system based in Bayes Decision Rule. They proposed to maximize the expression

$$J(C_{1,n}) = argmax_{j_{1,n}} P(j_{1,n}|C_{1,n})$$

where $J(C_{1,n})$ is the sequence of n junctures. These junctures can be breaks or not breaks. C_i is the context information of the juncture, which considers two previous POS tags and the following to the position of the phrase boundary.

$P(j_{1,n}|C_{1,n})$ is calculated as

$$P(j_{1,n}|C_{1,n}) = \prod_{i=1}^{n} \frac{P(j_i|C_i)}{P(j_i)} P(j_i|j_{i-l} \cdots j_{i-1})$$

where $P(j_i|C_i)$ is the probability of a juncture according to the adjacent tags, $P(j_i)$ is the probability of each juncture (break or non-break), and $P(j_i|j_{i-l} \cdots j_{i-1})$ is the n-gram of the juncture probability according to the previous l junctures.

Sun et al. [4] extended the approach of Black and Taylor estimating the probabilities $P(j_i|C_i)$ using binary decision trees.

In summary, there are several data-driven methodologies that achieve good results. However, most of the experiments have been done in English with different data, which turns difficult to make a fair comparison.

In this paper our baseline system consists in a finite state transducer that predicts phrase breaks given part-of-speech tags, as shown in next section. Although the algorithm can be used to predict phrase breaks in general (with or without a pause), it is only applied to phrase breaks with pauses.

2.1. Phrase break prediction using FST

The baseline system consists in a transducer that performs the conversion of part-of-speech tags into phrase break boundary tags. In the training step, the transducer is given a sequence of pairs of part-of-speech - phrase break boundary tags:

$$(p_1, b_1)(p_2, b_2)...(p_n, b_n) \quad (1)$$

where p_i is the part-of-speech tag of word w_i, and b_i indicates the existence of a phrase break boundary tag (B) or not ($\neg B$) after the word w_i.

The task of the transducer is to find the sequence of phrase break boundary tags that maximize the equation 2.

$$argmax_b P(b/p) = argmax_b \frac{P(b,p)}{P(p)} = argmax_b P(b,p) \quad (2)$$

$P(b,p)$ is the joint probability of a sequence of part-of-speech and phrase break boundary tags. This can be modeled using n-grams, as shown in equation 3.

$$P(b,p) = \prod_{i=1}^{N} P(b_i, p_i/b_{i-k}^{i-1}, p_{i-k}^{i-1}) \quad (3)$$

The language model obtained with the n-grams can be represented as a finite state automata (FSA). Each state represents a history $(b_{i-k}^{i}, p_{i-k}^{i})$ and the arcs contain the conditional probability of an observation given the previous history

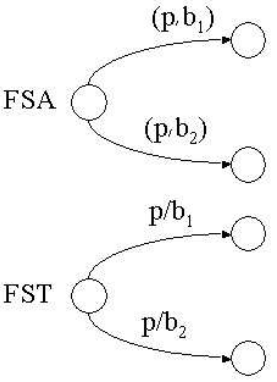


Figure 2: *FSA and FST.*

$(P(b_i, p_i/b_{i-k}^{i-1}, p_{i-k}^{i-1}))$. In this way, the joint probability of a sequence of observations can be obtained travelling the finite state automata given the observations, as shown in equation 4.

$$P(b,p) = P(b_1, p_1).P(b_2, p_2/b_1, p_1).$$
$$.P(b_3, p_3/(b_1, p_1)(b_2, p_2))... \quad (4)$$

In this paper, n-grams are estimated using variable length n-grams [5].

The FSA is converted into a FST, taking into account that the observation of a part-of-speech p_i produces an output b_i, as shown in Figure 2. Given the inputs p_i, there are several possible paths in the FST that can be travelled with the sequence p. Viterbi decoding is used to obtain the path that maximizes $P(b/p)$. Given the optimal state sequence, it is posible to obtain the phrase break boundary tags (b_i) that correspond to the best path through the FST.

FST's have been used in several tasks, such as phonetic transcription [6] and machine language translation [7]. These tasks are more complex, because in some cases there is a mapping of many-to-many from input to output. In addition, in some cases the output sequence has a different order than the input.

In this approach we decided to use part-of-speech as input due to two reasons:

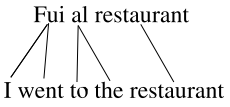
- **Reduction of the size of the input space**. The part-of-speech tags are used instead of words. The use of words would cause a need of a huge amount of corpus in order to obtain reliable probability estimations.
- **Relationship between part-of-speech tags and phrase breaks**. Several works in the area have shown that part-of-speech tags are an important source of information to decide the placement of a phrase break boundary [1, 3].

3. Proposed approach

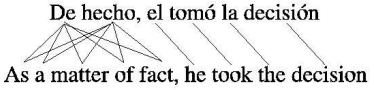
In the framework of speech-to-speech translation is encouraging the use of the pauses of the input speaker. Imitating the source speaker together with the use of voice conversion techniques may increase the resemblance of the synthetic output with the source speaker. What is more, such pauses will keep the discourse style and the intented meaning of the speaker avoiding misinterpretations or ambiguities introduced by prediction techniques explained in Section 2.

The SLT component provides the source language text and its translation into the target language with additional information about the cross-lingual correspondence of the words. Many situations arise with these links:

- **One-to-many.** Words in the source language may be translated into one or more words into the target language, because of lexical, grammatical, syntactical or semantical reasons. For example, in the sentence ”Yo fui al restaurant” (I went to the restaurant), ”al” is aligned with ”to the”. The word ”al” is a contraction of the Spanish words ”a el”.



- **Many-to-many.** Some constructions in the source language must be aligned with its counterpart in the target language in order to preserve the full meaning. For example, in the sentence ”De hecho, él tomó la decisión.” (As a matter of fact, he took the decision.), the words ”de hecho” are aligned with ”as a matter of fact”. Both sequences of words are constructions of Spanish and English that can not be splitted into a smaller unit without losing its meaning.



- **No alignment.** In some situations a given word in a language can not be aligned. These words are only used for particular reasons and do not have a counterpart in the other language.

In our proposal we make use of the tuples that can be formally defined as the set of shortest phrases that provides a monotonic segmentation of the bilingual data. An example is shown in Figure 3.

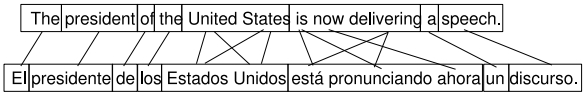


Figure 3: Example of a segmentation using tuples.

The analysis of a bilingual corpus with pauses in the source language has shown that most of the pauses in the boundary of a tuple have a corresponding pause in the target language. In this paper we propose to use this fact to transfer source pauses into the target language.

One of the important consequences of using tuples is the avoidance of the reordering effects in pause transfer. For example, in the sentence ”The White House PAUSE”, the pause after House will be transferred to ”La Casa PAUSE Blanca” in case of using words instead of tuples. The pause is inside the noun phrase. Meanwhile, tuples provide a right pause transfer, because ”White House” and ”Casa Blanca” is a tuple, and therefore the pause is transferred to the boundary of the tuple: ”La Casa Blanca PAUSE”.

However, a limitation arise when a pause falls inside a group of words with many-to-many aligments. In this case it is not posible to accurately find the position of the pause in the target language. Many words in the target language have a link with the word previous to the pause in the source language. The same happens when an alignment is missing.

The missing pauses due to the previously mentioned limitation are predicted using the approach proposed in Section 2. The algorithm will be used to predict the missing pauses, keeping in mind that some pauses are already placed. This will change the optimal output of the search algorithm compared to the baseline approach.

In the next section we show experimental results that support this approach.

4. Experiments

4.1. Experimental conditions

In these experiments we use a corpus of the TC-STAR project corresponding to a biligual male speaker (British English and Spanish). The pauses are automatically detected and annotated using a speech recognizer. The corpus consists of 197 paragraphs of the Parliamentary domain uttered by a bilingual speaker using a parliamentary style.

In these experiments we made the assumption that the translation is perfect. We use a bilingual corpus performed by human translators. The only errors in the corpus are in the alignment (generated using GIZA++ [8]) and in the detection of pauses (using the automatic speech recognizer named RAMSES [9]).

The corpus is divided in ten parts to perform 10-Fold experiments. Several metrics are used to study the performance of the systems: precision, recall and F-measure.

Two experiments will be conducted. In the first experiment we use as reference the pauses of the speaker in the target language.

Although the speaker was instructed to follow the same style in both languages, some pauses in one language are missing in the other, and viceversa. Due to this, in the second experiment each pause of the proposed technique will be analyzed to study its feasibility.

4.2. Experimental results

The experimental results of the baseline technique are shown in Table 1. Table 2 shows the results for the proposal. The proposed technique has a lower precision than the baseline due to a higher number of predicted pauses (a 20% of additional pauses). However, recall measures show that predicted pauses with the proposed method are better than the baseline technique.

Precision	Recall	F-Measure
69.23	68.18	68.70
58.00	66.41	61.92
50.68	61.15	55.43
54.77	58.10	56.39
57.66	58.95	58.30
56.36	72.65	63.48
55.74	67.83	61.19
55.68	67.39	60.98
61.78	62.58	62.17
59.54	69.64	64.19

Table 1: Experimental results for baseline approach.

Precision	Recall	F-Measure
66.17	68.18	67.16
58.38	71.75	64.38
49.34	61.98	54.94
54.43	62.16	58.04
56.08	61.94	58.86
56.39	75.78	64.66
52.97	68.53	59.75
51.39	66.66	58.04
60.23	66.45	63.19
58.15	73.21	64.82

Table 2: Experimental results for proposed approach.

As explained before, in some cases the pauses in the source language are not placed in the corresponding position in the target language. The bilingual speaker made a different choice in the target language utterance. In Table 3 we show the experimental results of the analysis of the predicted pauses for the proposed technique. Every pause is studied to decide manually its correctness.

These results show the real performance of the proposed approach avoiding the inconsistencies in the bilingual database.

Precision	Recall	F-Measure
83.73	75.95	79.65
75.00	80.23	77.52
71.03	76.92	73.86
75.87	76.26	76.07
77.09	75.00	76.03
76.68	83.61	80.00
72.03	77.55	74.69
73.76	78.01	75.82
78.60	75.59	77.07
75.90	81.81	78.75

Table 3: Experimental results for proposed approach after manual supervision.

In the next two paragraphs we show an example of a parallel output of the system. Source language is English and target language is Spanish. Source paragraph has detected pauses from the speaker using ASR, while target paragraph has predicted pauses. Those pauses predicted using alignment information and tuples are stated as **(PAUSE)**, while pauses predicted using FST are indicated as **PAUSE**.

SOURCE (English): So all of the commissioners (PAUSE) nominated in November can expect tough questioning from MEPS (PAUSE) of the political centre, (PAUSE) the kind of questioning that was too little in evidence (PAUSE) at last month's hearings. (PAUSE) this commision is the first (PAUSE) of a new enlarged Europe (PAUSE) and the imperative for liberals (PAUSE) in this house is to ensure (PAUSE) that it is effective, (PAUSE) committed and competent. (PAUSE)

TARGET (Spanish): Así pues, PAUSE todos los comisarios (PAUSE) nombrados en noviembre pueden esperar un buen interrogatorio PAUSE por parte de los diputados (PAUSE) del centro político, (PAUSE) un tipo de interrogatorio PAUSE demasiado poco presente (PAUSE) en las audiencias del mes pasado. (PAUSE) Esta comisión PAUSE es la primera (PAUSE) de una nueva Europa ampliada (PAUSE) y el imperativo de los liberales (PAUSE) de esta cámara es garantizar (PAUSE) que sea eficaz, (PAUSE) comprometida PAUSE y competente. (PAUSE)

In this example, the eleven pauses of source language are correctly transfered using tuples. Here we can observe how powerful this technique can be imitating the source speaker.

5. Conclusions

In this paper we presented two techniques to predict pauses for text-to-speech in the speech-to-speech translation framework and the parliamentary domain.

In the first case we use a finite state transducer. The poor prediction results indicate that the pauses are not located in the same place than in the source language, as desired.

In the second case we use a combination of a finite state transducer and a pause transfer algorithm that takes advantage of alignment information provided by SLT. The better results show that many pauses are placed as expected, imitating the source speaker style.

The results are encouraging to continue the research in this topic by studying the problem of the pauses inside tuples.

6. References

[1] Prieto, P.; Hirschberg, J., 1996. Training intonational phrasing rules automatically for English and Spanish text-to-speech. *Speech Communication*, 18, 155-158.

[2] Koehn, P.; Abney, S.; Hirschberg, J.; Collins, M., 2000. Improving intonational phrasing with syntactic information. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 3, 1289-1292.

[3] Black, A.; Taylor, P., 1997. Assigning phrase breaks from part-of-speech sequences. *Proceedings of European Conference on Speech Communication and Technology*, 995-998.

[4] Sun, X.; Applebaum, T.H., 2000. Intonational phrase break prediction using decision tree and n-gram model. *Proceedings of European Conference on Speech Communication and Technology*, 537-540.

[5] Bonafonte, A., 1996. Language modeling using x-grams. *Proceedings of International Conference on Spoken Language Processing*, 394-397.

[6] Galescu, L.; Allen, J.F., 2001. Bi-directional conversion between graphemes and phonemes using a joint n-gram model. *Proceedings of the 4th ISCA workshop on Speech Synthesis*, 103-108.

[7] de Gispert, A.; Mariño, J.B., 2002. Using x-grams for speech-to-speech translation. *Proceedings of International Conference on Spoken Language Processing*, 1885-1888.

[8] Och, F.J.; Ney, H., 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), 19-51.

[9] Bonafonte, A.; Mariño, J.B.;Nogueiras, A.;Fonollosa, J.A.R., 1998. RAMSES: el sistema de reconocimiento del habla continua y gran vocabulario desarrollado por la UPC. *VIII Jornadas de Telecom I+D (TELECOM I+D'98)*.