Automatic Extraction of Parallel Speech Corpora from Dubbed Movies

Alp Öktem[1] and **Mireia Farrús**[1] and **Leo Wanner**[2,1]
[1]Universitat Pompeu Fabra, Spain
[2]Catalan Institute for Research and Advanced Studies (ICREA), Spain

Abstract

This paper presents a methodology to extract parallel speech corpora based on any language pair from dubbed movies, together with an application framework in which some corresponding prosodic parameters are extracted. The obtained parallel corpora are especially suitable for speech-to-speech translation applications when a prosody transfer between source and target languages is desired.

1 Introduction

The availability of large parallel corpora is one of the major challenges in developing translation systems. Bilingual corpora, which are needed to train statistical translation models, are harder to acquire than monolingual corpora since they presuppose the implication of labour in translation or interpretation. Working in the speech domain introduces even more difficulties since interpretations are not sufficient to capture the paralinguistic aspects of speech. Several attempts have been recently made to acquire spoken parallel corpora of considerable size. However, these corpora either do not reflect the prosodic aspects in the interpreted speech or do not carry the traits of natural speech. Or they simply do not align well the source and the target language sides.

To account for this deficit, we propose to exploit dubbed movies where expressive speech is readily available in multiple languages and their corresponding aligned scripts are easily accessible through subtitles. Movies and TV shows have been a good resource for collecting parallel bilingual data because of the availability and open access of subtitles in different languages. With 1850 bitexts of 65 languages, the OpenSubtitles project (Lison and Tiedemann, 2016) is the largest resource of translated movie subtitles compiled so far. The time information in subtitles makes it easy to align sentences of different languages since timing is correlated to the same audio (Itamar and Itai, 2008). In the presence of multiple aligned audio for the same movie, the alignment can be extended to obtain parallel speech corpora. Popular movies, TV shows and documentaries are released with dubbed audio in many countries. Dubbing requires the voice acting of the original speech in another language. Because of this, the dubbed speech carries more or less the same paralinguistic aspects of the original speech.

In what follows, we describe our methodology for the extraction of a speech parallel corpus based on any language pair from dubbed movies. Unlike Tsiartas et al. (2011), who propose a method based on machine learning for automatically extracting bilingual audio-subtitle pairs from movies, we only need raw movie data, and do not require any training. Moreover, our methodology ensures the fulfilment of the following requirements: (a) it is easily expandable, (b) it supports multiple pairs of languages, (c) it can handle any domain and speech style, and (d) it delivers a parallel spoken language corpus with annotated expressive speech. "Expressive speech" annotation means that the corpus is prosodically rich, which is essential to be able to deal with non-neutral speech emotions, as done in increasingly popular speech-to-speech translation applications that try to cope with prosody transfer between source and target utterances (Agüero et al., 2006; Sridhar et al., 2008; Anumanchipalli et al., 2012).

The remainder of the paper is structured as follows. Section 2 reviews the main multilingual parallel speech corpora available to the research community. Section 3 presents the methodology used in the current paper, and Section 4 discusses the current state of the obtained parallel corpora so far.

Proceedings of the 10th Workshop on Building and Using Comparable Corpora, pages 31–35,
Vancouver, Canada, August 3, 2017.

In Section 5, finally, some conclusions are drawn and some aspects of our future work in the context of parallel speech corpora are mentioned.

2 Available Parallel Speech Corpora

As already mentioned above, several attempts have been made to compile large spoken parallel corpora. Such corpora of considerable size are, e.g., the EPIC corpus (Bendazzoli and Sandrelli, 2005), the EMIME Bilingual Database (Wester, 2010), and the Microsoft Speech Language Translation (MSLT) corpus (Federmann and Lewis, 2016). All of them have been manually compiled, and all of them show one or several shortcomings. The EPIC corpus, which has been compiled from speeches from the European Parliament and their interpretations, falls short in reflecting the prosodic aspects in the interpreted speech. The EMIME database is a compilation of prompted speeches and does not capture the natural spoken language traits. The MSLT corpus has been collected in bilingual conversation settings, but there is no one-to-one alignment between sentences in different languages. A summary of the available bilingual speech corpora is listed in Table 1.

3 Methodology

Our multimodal parallel corpus creation consists of three main stages: (1) movie sentence segmentation, (2) prosodic parameter extraction, and (3) parallel sentence alignment. The first and second stages can be seen as a monolingual data creation, as they take the audio and subtitle pairs as input in one language, and output speech/text/prosodic parameters at the sentence level. The resulting monolingual data from stages 1 and 2 are fed into stage 3, where corresponding sentences are aligned and reordered to create the corresponding parallel data. A general overview of the system is presented in Figure 1.

Let us discuss each of these stages in turn.

3.1 Segmentation of movie audio into sentences

This stage involves the extraction of audio and complete sentences from the original audio and the corresponding subtitles of the movie. For subtitles, the SubRip text file format[1] (SRT) is accepted. Each subtitle entry contains the following information: (i) start time, (ii) end time, and (iii) text of the speech spoken at that time in the movie. The subtitle entries do not necessarily correspond to sentences: a subtitle entry may include more than one sentence, and a sentence can spread over many subtitle entries; consider an example portion of a subtitle:

```
80
00:06:46,114 --> 00:06:48,741
Well, I was stationed
up in Casablanca

81
00:06:48,825 --> 00:06:51,535
at an army field hospital
during the war.

82
00:06:51,995 --> 00:06:53,871
- Do you live in Morocco?
- Yes.
```

The sentence segmentation stage starts with a preprocessing step in which elements that do not correspond to speech are removed. These include: Speaker name markers (e.g., JAMES: ...), text formatting tags, non-verbal information (laughter, horn, etc.) and speech dashes. Audio is initially segmented according to the timestamps in subtitle entries, with extra 0.5 seconds at each end. Then, each audio segment and its respective subtitle text are sent to the speech aligner software (Vocapia Scribe[2]) to detect word boundaries. This pre-segmentation helps to detect the times of the words that end with a sentence-ending punctuation mark ('.', '?', '!', ':', '...'). Average word boundary confidence score of the word alignment is used to determine whether the sentence will be extracted successfully or not. If the confidence score is above a threshold of 0.5, the initial segment is cut from occurrences of sentence-endings. In a second pass, cut segments that do not end with a sentence-ending punctuation mark are merged with the subsequent segments to form full sentences. We used *Libav*[3] library to perform the audio cuts.

3.2 Prosodic parameter extraction

This stage involves prosodic parameter extraction for each sentence segment detected in stage 1. The *ProsodyPro* library (Xu, 2013) (a script developed for the Praat software (Boersma and Weenink, 2001)) is used to extract prosodic features from speech. As input, ProsodyPro takes the audio of

[1] https://www.matroska.org/technical/specs/subtitles/srt.html

[2] https://scribe.vocapia.com/

[3] https://libav.org/

Corpus	Languages	Speech style
EPIC	English, Italian, Spanish	spontaneous/interpreted
MSLT	English, French, German	constrained conversations
EMIME	Finnish/English, German/English	prompted
EMIME Mandarin	Mandarin/English	prompted
MDA (Almeman et al., 2013)	Four Arabic dialects	prompted
Farsi-English (Melvin et al., 2004)	Farsi/English	read/semi-spontaneous

Table 1: Some available parallel speech corpora.

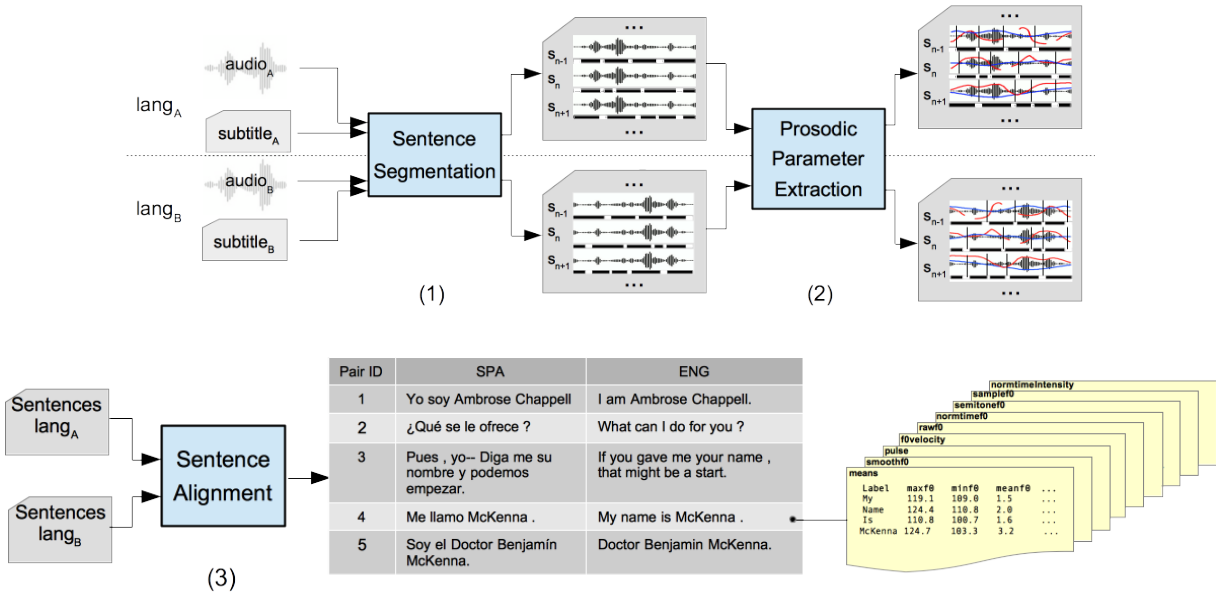

Figure 1: Above: Monolingual corpus creation from different audio-subtitle pairs in parallel. Below: Bilingual parallel corpus creation of the example dataset.

an utterance and a TextGrid file containing word boundaries and outputs a set of objective measurements suitable for statistical analysis. We run ProsodyPro for each audio and TextGrid pair of sentences to generate the prosodic analysis files. See Table 2 for the list of analyses performed by ProsodyPro (Information taken from ProsodyPro webpage[4]).

The TextGrid file with word boundaries is produced by sending the sentence audio and transcript to the word-aligner software and then converting the alignment information in XML into TextGrid format. Having word boundaries makes it possible to align continuous prosodic parameters (such as pitch contour) with the words in the sentence.

3.3 Parallel sentence alignment

This stage involves the creation of the parallel data from two monolingual data obtained from different audio and subtitle pairs of the same movie. The goal is to find the corresponding sentence s_2 in language 2, given a sentence s_1 in language 1. For each s_1 with timestamps (s_{s_1}, e_{s_1}), s_2 is searched within a sliding window among sentences that start in the time interval [s_{s_1} - 5, s_{s_1} + 5]. Among candidate sentences within the range, the most similar to s_1 is found by first translating s_1 to language 2 and then choosing the $\{s_1, s_2\}$ pair that gives the best translation similarity measure above a certain threshold. For translation, the *Yandex Translate* API[5] and for similarity measure the *Meteor* library (Denkowski and Lavie, 2014) is used.

4 Obtained Corpus and Discussion

We have tested our methodology on three movies, which we retrieved from the University Library: *The Man Who Knew Too Much* (1956), *Slow West* (2015) and *The Perfect Guy* (2015). The movies are originally in English, but also have dubbed Spanish audio. English and Spanish subtitles were

[4] http://www.homepages.ucl.ac.uk/~uclyyix/ProsodyPro/

[5] https://tech.yandex.com/translate/

ProsodyPro output file	Description
rawf0	Raw f0 contour in Hz
f0	Smoothed f0 with trimming algorithm (Hz)
smoothf0	Smoothed f0 with triangular window (Hz)
semitonef0	f0 contour in semitones
samplef0	f0 values at fixed time intervals (Hz)
f0velocity	First derivative of f0
means	f0, intensity and velocity parameters (mean, max, min) for each word
normtimef0	Constant number of f0 values for each word
normtimeIntensity	Constant number of intensity values for each word

Table 2: Some of the files generated by ProsodyPro.

acquired from the *opensubtitles* webpage[6].

At the time of the submission, we have automatically extracted 2603 sentences in English and 1963 sentences in Spanish summing up to 80 and 49 minutes of audio respectively and annotated with prosodic parameters. 1328 of these sentences were aligned to create our current parallel bilingual corpora. We are in the process of expanding our dataset.

Due to the copyright on the movies, we are unable to distribute the corpus that we extracted. However, using our software, it is easy for any researcher to compile a corpus on their own. For testing purposes, English and Spanish subtitles and audio of a small portion of the movie *The Man Who Knew Too Much*, as well as the parallel data extracted with this methodology are made available on the github page of the project.

Movie ID	# sentences extracted (eng / spa)	# sentences aligned (parallel)
slow.west	414 / 315	237
tmwktm	1429 / 813	599
perfect.guy	760 / 835	492
TOTAL	2603 / 1963	1328

Table 3: Process results for three movies.

Lang.	# subtitle entries	# sentence end marks	# sentences extracted
eng	1743	1681	1429
spa	1266	1613	813

Table 4: Sentence extraction statistics in English (original audio) and Spanish (dubbed audio) of the movie *The Man Who Knew Too Much*.

Table 3 lists the number of monolingual and parallel sentences obtained from the three movies so far. We observe that the number of Spanish sentences extracted in stage 2 is sometimes lower than the number of English sentences. This is mainly because of the translation difference between the Spanish subtitles and the dubbed Spanish audio. Subtitles in languages other than the original language of the movie do not always correspond with the transcript used in dubbing. If the audio and the text obtained from the subtitle do not match, the word aligner software performs poorly and that sentence is skipped. This results in fewer number of extracted sentences in dubbed languages of the movie. Table 4 shows more in detail the effect of this. Poor audio-text alignment results in loss of 15.0% of the sentences in original audio, whereas in dubbed audio this loss increases to 49.6%.

Another major effect on detection of sentences is the background noise. This again interferes with the performance of the word aligner software. But since samples with less background noise is desired for a speech database, elimination of these samples is not considered as a problem.

5 Conclusions and Future Work

We have presented a methodology for the extraction of multimodal speech, text and prosody parallel corpora from dubbed movies. Movies contain large samples of conversational speech, which makes the obtained corpus especially useful for speech-to-speech translation applications. It is also useful for other research fields such as large comparative linguistic and prosodic studies.

As long as we have access to a matching pair of audio and subtitles of movies, the corpora obtained can be extended as a multilingual speech parallel corpora adaptable to any language pair. Moreover, it is an open-source tool and it can be

[6] https://www.opensubtitles.org/

adapted to any other prosodic feature extraction module in order to obtain a customized prosody parallel corpus for any specific application. The code to extract multilingual parallel corpora together with a processed sample movie excerpt is open source and available to use[7] under the GNU General Public License[8].

As future work, we plan to extend our corpus in size and make the parallel prosodic parameters available online. We also plan to replace the proprietary word aligner tool we are using with an open source alternative with better precision and speed.

Acknowledgments

We would like to thank Alicia Burga for giving the initial idea of this work. This work is part of the KRISTINA project, which has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under the Grant Agreement number 645012. The second author is partially funded by the Spanish Ministry of Economy, Industry and Competitiveness through the *Ramón y Cajal* program.

References

Pablo D. Agüero, Jordi Adell, and Antonio Bonafonte. 2006. Prosody generation for speech-to-speech translation. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, volume 1, pages 557–560.

Khalid Almeman, Mark Lee, and Ali Abdulrahman Almiman. 2013. Multi dialect Arabic speech parallel corpora. In *1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*. IEEE, pages 1–6.

Gopala Krishna Anumanchipalli, Luís C. Oliveira, and Alan W. Black. 2012. Intent transfer in speech-to-speech machine translation. In *Spoken Language Technology (SLT) Workshop*. IEEE, pages 153–158.

Claudio Bendazzoli and Annalisa Sandrelli. 2005. An approach to corpus-based interpreting studies: developing EPIC (European Parliament Interpreting Corpus). In *MuTra 2005–Challenges of Multidimensional Translation*. pages 1–12.

Paul Boersma and David Weenink. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5(9/10):341–345.

[7] https://github.com/TalnUPF/movie2parallelDB

[8] http://www.gnu.org/licenses

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL Workshop on Statistical Machine Translation*.

Christian Federmann and William D. Lewis. 2016. Microsoft Speech Language Translation (MSLT) corpus: The IWSLT 2016 release for English, French and German. In *International Workshop on Spoken Language Translation*.

Einav Itamar and Alon Itai. 2008. Using movie subtitles for creating a large-scale bilingual corpora. In *6th International Conference on Language Resources and Evaluation (LREC)*.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*. pages 923–929.

Robert S. Melvin, Win May, Shrikanth S. Narayanan, Panayiotis G. Georgiou, and Shadi Ganjavi. 2004. Creation of a doctor-patient dialogue corpus using standardized patients. In *4th International Conference on Language Resources and Evaluation (LREC)*.

Vivek Kumar Rangarajan Sridhar, Srinivas Bangalore, and Shrikanth S. Narayanan. 2008. Factored translation models for enriching spoken language translation with prosody. In *Interspeech*. pages 2723–2726.

Andreas Tsiartas, Prasanta Ghosh, Panayiotis G Georgiou, and Shrikanth Narayanan. 2011. Bilingual audio-subtitle extraction using automatic segmentation of movie audio. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pages 5624–5627.

Mirjam Wester. 2010. The EMIME Bilingual Database. Technical report, The University of Edinburgh.

Yi Xu. 2013. ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis. In *Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*. pages 7–10.