Lang Resources & Evaluation (2021) 55:925–946
https://doi.org/10.1007/s10579-021-09556-2

ORIGINAL PAPER

Corpora compilation for prosody-informed speech processing

Alp Öktem[1] · **Mireia Farrús**[2] · **Antonio Bonafonte**[3]

Accepted: 29 July 2021 / Published online: 4 September 2021

Abstract Research on speech technologies necessitates spoken data, which is usually obtained through read recorded speech, and specifically adapted to the research needs. When the aim is to deal with the prosody involved in speech, the available data must reflect natural and conversational speech, which is usually costly and difficult to get. This paper presents a machine learning-oriented toolkit for collecting, handling, and visualization of speech data, using prosodic heuristic. We present two corpora resulting from these methodologies: PANTED corpus, containing 250 h of English speech from TED Talks, and Heroes corpus containing 8 h of parallel English and Spanish movie speech. We demonstrate their use in two deep learning-based applications: punctuation restoration and machine translation. The presented corpora are freely available to the research community.

Keywords Speech corpus · Parallel data · Speech transcription · Spoken machine translation · Punctuation · Pause · F0 · Intensity

✉ Alp Öktem
alp@collectivat.cat

Mireia Farrús
mfarrus@ub.edu

Antonio Bonafonte
antonio.bonafonte@upc.edu

[1] Universitat Pompeu Fabra/Col·lectivaT, Barcelona, Spain

[2] Universitat Pompeu Fabra/Universitat de Barcelona, Barcelona, Spain

[3] Universitat Politècnica de Catalunya, Barcelona, Spain

1 Introduction

Prosody is an important aspect in human spoken communication (Fujisaki, 1997), as it acts as an essential element from pointing out relevance on certain aspects to expressing emotions or to help structure discourse. However, despite its crucial role in speech, prosody is still a less-focused phenomena in systems that process spoken language. This is mainly due to its additional complexity. Prosodic features have continuous representations, span over multiple segments in speech—therefore known as *suprasegmental* features, and their interpretation depends on the context (Crystal, 2003; Fujisaki, 1997). For this reason, there is a current tendency in these systems to carry on with linguistic modeling once spoken input is converted to its written form. Although conversion to discrete linguistic units facilitates computational processing, modeling only lexical information in speech leaves out any important aspect of it carried through prosody.

In this paper, we deal with the compilation of speech corpora annotated with prosody heuristic, adapted for machine-learning based applications. We propose a minimal spoken language data structure that facilitates joint processing of prosodic and other linguistic features in speech. We furthermore present a set of tools for harvesting, compiling and visualizing this type of data. Our strategy for speech data collection is based on exploiting readily available recorded material—usually referred to as *found data*. The use of found data requires some data pre-processing in order to convert raw speech material into sampled and annotated speech. Our methodologies built on our previous work (Öktem, Farrús, & Bonafonte, 2018) facilitate the collection of both monolingual and parallel spontaneous speech data in a scalable manner.

The proposed corpus compilation methodology is put into use in the collection of two datasets: an English conference speech corpus and an English–Spanish dubbed movie speech corpus. All the developed methodologies and corpora are made publicly available as open source software libraries[1] and through the Pompeu Fabra University (UPF) Digital Repository[2], respectively.

The contributions of this paper, and their corresponding sections are listed in Table 1. Section 5 presents a brief evaluation of both corpora as source data for speech recognition and translation tasks, and finally Sect. 6 concludes the paper.

2 Related work

Computational applications that deal with prosody necessitate a standard for representing the data structure; i.e. the structure of speech with its orthography and prosody together. One of the most popular of these conventions is the *TextGrid* format, which is used by *Praat* (Boersma & Weenink, 2019). A TextGrid file stores any number of tiers that can be used to label acoustic and linguistic information time-aligned with speech. Although very useful for visualization in Praat, this

[1] Available in http://www.github.com/alpoktem.

[2] http://repositori.upf.edu.

Table 1 Paper contributions

Name	Description	Section
Proscript	Prosody extraction software based on a minimal data-structure	3.1
Prosograph	Software for speech data visualization	3.2
movie2parallelDB	Software for monolingual and parallel spoken data collection	3.3
PANTED corpus	250 h speech corpus from TED talks	4.1
Heroes corpus	Parallel English–Spanish speech corpus of dubbed movie segments	4.2

format is not designed to be functional for viewing and manipulating computationally by itself. Every tier defines which event occurs at what time on its own and it is difficult to associate events that occur in parallel in different tiers. Also, for storage of raw acoustic features, Praat uses different file formats. Due to this design, a complete prosodic–acoustic representation of a short utterance ends up being represented with a clutter of files. Other tools such as (Huang, Chen, & Harper, 2006; Xu, 2013) are also based on Praat and are only runnable through its interface. And some others such as AuToBI (Rosenberg, 2010) or ANALOR (Avanzi, Lacheret-Dujour, & Victorri, 2008) provide automatic or semi-automatic annotation of English and French prosodic structures, respectively, while ELAN (Sloetjes & Wittenburg, 2008) serves as an annotation tool for audio and video, compatible with Praat TextGrid format.

Related to securing spoken parallel corpora, several attempts have been carried out, as for instance: the European Parliament Interpretation Corpus (EPIC) corpus (Bendazzoli & Sandrelli, 2005), the EMIME (Effective Multilingual Interaction in Mobile Environments) Bilingual prompted speech (Wester, 2010), the Microsoft Speech Language Translation (MSLT) corpus (Federmann & Lewis, 2016), or the Multi Dialect Arabic (MDA) parallel prompted speech corpora (Almeman, Lee, & Almiman, 2013). The *300 Languages Project*, part of the Rosetta Project,[3] aims at collecting parallel audio and texts in the world's 300 most widely-spoken languages. More recently, the CPJD corpus (Takamichi & Saruwatari, 2018) provides crowd-sourced parallel speech data of Japanese dialects. In addition, MaSS corpus is a multilingual parallel speech dataset based on recorded readings of the Bible with speech-to-text and speech-to-speech alignments (Zanon Boito et al., 2020).

The need for monolingual spoken data is growing steadily to achieve linguistic coverage in automatic speech recognition and text-to-speech research and development. Some examples to these are: the Switchboard corpus (Godfrey & Holliman, 1993), for English telephone conversational speech, the CALLHOME speech corpora, consisting of telephone conversations in several languages (Canavan, Graff, & Zipperlen, 1997), English Boston University Radio Speech Corpus (Ostendorf, Price, & Shattuck-Hufnagel, 1996), Rhapsodie (Lacheret et al., 2014), a French speech corpus with prosodic, syntactic and orthographic annotations, DEMoS (Parada-Cabaleiro et al., 2019) an Italian emotional speech corpus, RSC

[3] https://rosettaproject.org/projects/300-languages/.

(Georgescu et al., 2020), a Romanian read speech corpus for automatic speech recognition, TV3Parla (Külebi & Öktem, 2018) and ParlamentParla (Külebi et al., 2020), parliamentary and television speech corpora for Catalan.

The vast amount of natural language data residing in the web has been an invaluable source for both linguistic research and language technology development. OPUS collection (Tiedemann, 2012), for instance, is a publicly available corpus of parallel text from the web. Among others, it includes the *OpenSubtitles* collection (Lison & Tiedemann, 2016; Lison et al., 2018) of translated movie subtitles. Fortunately, although not created for research purposes and although being monolingual, TED talks[4] have become an inestimable large and free resource for research in spoken language. The number or works that have already used these resources is relevant (Cettolo, Girardi, & Federico, 2012; Pappas & Popescu-Belis, 2013; Farrús, Lai, & Moore, 2016). However, since TED talks were not thought to fit computational linguistics research objectives, they require a significant work on pre-processing the data. Some attempts for such processing was made for TED-LIUM corpus (Rousseau, Deléglise, & Estève, 2012), an English speech recognition training corpus from TED talks, and also for text-based tasks, such as document classification (Hermann & Blunsom, 2014) and machine translation (Cettolo et al., 2012). One of the aims of this paper is to present a prepared dataset based on TED talks for prosody modeling in speech applications.

3 Toolkit for speech corpus creation

This section describes the main methodology and toolkit employed in the creation of our corpora:

(i) *Proscript* for joint lexical–prosodic data handling,
(ii) *Prosograph* for visualization of large speech corpora, and
(iii) *movie2parallelDB* for obtaining parallel speech corpora from dubbed media.

These tools were useful for the development and creation of our corpora to adapt existing formats, align text and speech content, and visualize prosodic phenomena.

3.1 Linguistically-aligned prosody extraction with *Proscript*

Prosody is conveyed through three different elements: intonation, stress, and rhythm. These elements are perceived by listeners as changes in fundamental frequency (F0), intensity and sound duration over time, respectively (Adami et al., 2003). At the same time, F0, intensity and duration can be extracted in terms of their acoustic form by means of speech processing tools. For the current work, we propose a minimal spoken language data structure for facilitating: (1) the creation and storage of spoken language data represented by these measurable features of

[4] https://www.ted.com/.

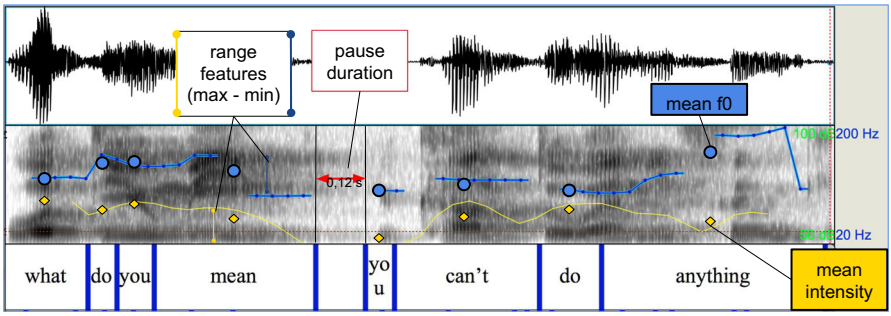

Fig. 1 Word-level prosodic feature labelling demonstrated on Praat

prosody—from now one referred to as *annotated* data, and (2) their processing with machine learning applications.

The idea behind our lexical–prosodic information structuring is based on representing relevant prosodic phenomena aligned with their respective linguistic context. Figure 1 illustrates this idea. In this example, the relevant features are pausing, intonation, intensity and inter-word prosodic changes. Instead of focusing on the complete movement of F0 or energy levels, average values are calculated within the period of utterance of a word. The same framework could be extended to represent contours as a vector. The aim is to represent the utterance in time-linearity that recurrent type neural networks are designed to process. Each segment is represented as a vector of linguistic and prosodic features. An example of a word-level segment where only mean values are calculated would be:

$$\begin{aligned} segment_i = \langle & word_i, mean_f0_i, range\_f0_i, \\ & mean\_int_i, range\_int_i, pause\_after_i \rangle. \end{aligned} \tag{1}$$

To address our general idea of representing linguistically aligned prosodic information, we have developed the *Proscript framework*. This framework provides a spoken language data representation format and a library for the creation, manipulation, reading and writing of this sort of data.

Proscript represents speech as features occurring in parallel at discrete bounded intervals, referred to as *segments*. Segments can be defined within, for instance, a word, a prosodic phrase, a sentence or a group of sentences. Any type of linguistic, prosodic or morphosyntactic features can be stored within these boundaries. Table 2 shows an example of parallel features stored in a Proscript file. Here, the linguistic units are words, and the set of features is determined by the application.

Proscript Python library[5] was developed to facilitate the creation, manipulation and annotation of Proscript files. It can be imported as a Python library to batch process transcribed speech files, annotate them with the desired features and output as files. Both word alignment and prosodic–acoustic tagging software (explained in following subsections) are accessible through the library. For word alignment, we

[5] http://github.com/alpoktem/proscript.

Table 2 An example list of features used in a Proscript format file with word-level segmentation

Feature	Details
Word	As a token
id	Unique word id
Speaker id	Unique speaker id
Start time	Start time of the word in an associated audio file
End time	End time of the word in an associated audio file
Pause	Coming before and after the word
Punctuation	Coming before and after the word
POS	Part-of-speech
ToBI	ToBI label
F0	Mean/max/min/std [in Hz and log-scaled (semitones)]
Intensity	Mean/max/min/std (in decibels and log-scaled)
F0 contour	As a list of values (semitones)
Intensity contour	As a list of values (log-scaled)
Speech rate	In second per syllable

used the open-source *Montreal Forced Aligner* (McAuliffe et al. 2013) for its availability of English and Spanish models. The forced alignment process is built on an automatic speech recognition system and requires its own acoustic models and a pronunciation dictionary. A Spanish pronunciation dictionary was created for the purpose of aligning Spanish language speech data using an open-source vocabulary[6] and *TransDic* phonetic transcription software (Garrido, Codina, & Fodge, 2018).[7] As for prosodic–acoustic feature annotation we used *ProsodyTagger* (Domínguez, Farrús, & Wanner, 2016a; Domínguez et al., 2016b), which is a Python wrapper around the feature extraction scripts from Praat.

In our deep learning-based applications, we opted for a log-scale representation for contour-style features with respect to the mean of the utterances. We computed the pitch range on a word to take into account the most common textual linguistic unit. Although speech is continuous, silences can abruptly change intonation range, and are still added between words, which reinforces the use of words as segments. Units that did not give any measurement were simply labeled as 0.0 to represent average of the speaker. The sampling frequency of our sounds is 16 kHz and the corresponding acoustic features were extracted at a time step of 10 ms.

3.2 Aiding study of large speech corpora with *Prosograph*

Being able to clearly visualize the different elements involved in prosody—intonation, rhythm, and stress—is often needed in computational prosody research. Several speech analysis tools (e.g. Praat), together with derived scripts and tools

[6] *ISpell*: https://www.gnu.org/software/ispell/.

[7] Resource available in: https://github.com/TalnUPF/phonetic_lexica.

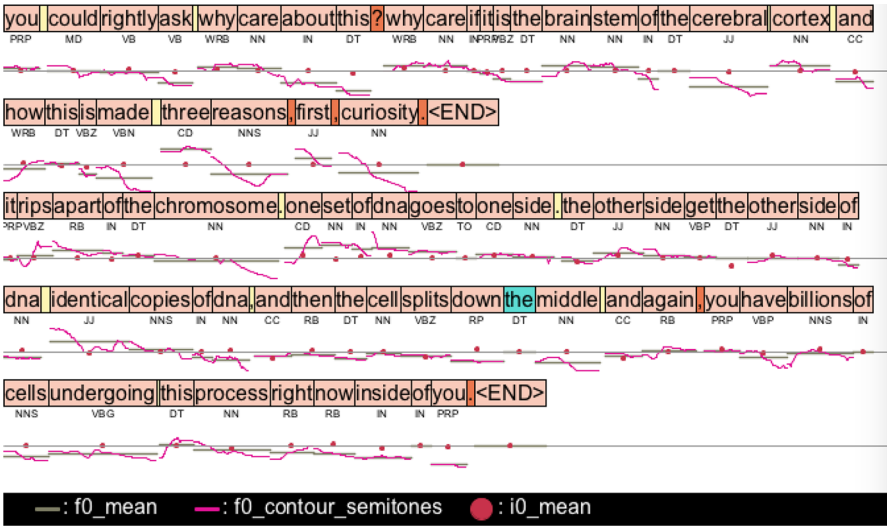

Fig. 2 An example of a visualization frame of segments from a conference talk with Prosograph. Two utterances are represented in the same view with readable transcription and traceable F0 contour

(Boersma & Weenink, 2019; Xu, 2013; Mertens, 2004; Domínguez et al., 2016b) partially cover these needs by helping to visualize quantifiable speech features like fundamental frequency (*F*0) and intensity contours, word stress marking, or prosodic labeling. These tools work well when showing detailed analyses on data and visualizing one single utterance at a time as in Fig. 1, but fail in visualizing generalized word-averaged speech features of many utterances, e.g., a discourse or a collection of speech samples, at once.

We developed *Prosograph* for visualizing acoustic and prosodic information of long speech segments together with their transcript. One of its principal motivations is to enable observing the relationship between prosodic features and punctuation in text. Moreover, its interactive interface makes it easy to listen to any portion of the displayed speech to accommodate auditory analysis (Öktem, Farrús, & Wanner, 2017b).

An overview can be seen in Fig. 2. Similar to sheet music representation, the prosodic feature values are plotted in the vertical axis over a temporal horizontal axis. Words are put in order together with pauses and punctuation, and the prosodic features are drawn under each corresponding word.

Prosograph can be customized easily as it is written in the highly visual and simple programming language *Processing*.[8] To demonstrate, we created a bilingual mode of Prosograph for analyzing parallel speech corpora. As illustrated in Fig. 3, the bilingual mode makes it possible to visualize aligned parallel corpora. Aligned samples are displayed side by side to accommodate e.g. prosodic comparison. Both

[8] http://processing.org.

Fig. 3 Visualization of parallel samples from an episode of Heroes corpus with bilingual mode of Prosograph

original[9] and bilingual mode[10] of Prosograph are made openly available as open-source software under the GNU General Public License.[11]

3.3 Automated speech corpus creation with *Movie2ParallelDB*

This section explains *movie2parallelDB*, our methodology to make parallel speech corpora from dubbed movies (Öktem et al., 2018). It makes only use of raw data: original and dubbed audio track of a multimedia such as a movie or television series together with their subtitles. It does not require any training as in the case of previous works (Tsiartas et al., 2011). We can list the main advantages of our methodology as follows:

(1) It is easily expandable to obtain monolingual or parallel corpora,
(2) It is language independent with sole dependence on availability of forced alignment models,
(3) It can handle any domain and speech style,
(4) It delivers a spoken language corpus with prosodic feature annotations, and
(5) It does not violate the fair use principles that go with copyrighted material.

[9] http://github.com/alpoktem/Prosograph.

[10] http://github.com/alpoktem/Prosograph2.

[11] http://www.gnu.org/licenses/gpl.html.

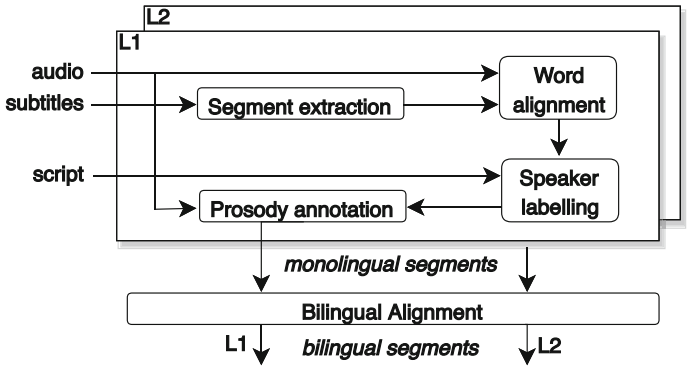

Fig. 4 Overall corpus extraction pipeline. Monolingual stage that involves segmentation, word alignment, speaker labelling and prosodic feature annotation is performed separately for both audio channels (L1 and L2). Alignment stage aligns segments of each language and outputs them as parallel segments

We publish movie2parallelDB as an open-source software together with instructions to use in http://github.com/alpoktem/movie2parallelDB.

3.3.1 Methodology

Figure 4 illustrates the overall process for mining speech samples from a dubbed multimedia. As raw data, we need audio tracks and subtitles in the original and dubbed languages, and a script that contains speaker information. In the first stage, we mine the speech segments at the sentence level using the audio and subtitles pair for each language. Later, we extract timing information at the word-level and obtain speaker labels for each segment from the script. Monolingual extraction stage is completed once we annotate the prosodic features using the Proscript library. Finally, we align the segments extracted in each language to obtain the parallel corpus segments.

For a detailed demonstration of the process of obtaining parallel segments from a portion of a movie, refer to Fig. 5. Subtitles are the source for obtaining both (1) audio transcriptions, and (2) timing information related to the utterances. These information are contained in a standard *srt* format[12] subtitles, entry by entry, where each subtitle entry is represented by an index, time cues and the script being spoken at that time in the movie. A subtitle entry can consist of a single, multiple (#3), or incomplete sentences (#1). They can contain speech from a single (#1, 2) or multiple speakers (#3). Thus, using only these time cues does not suffice for extracting audio segments with complete sentences of a single speaker. To achieve this, word boundary information is combined with punctuation mark positions to split and merge segments as needed. Two entries are merged if the first one does not end with a sentence-ending punctuation mark and the second one starts with a

[12] SubRip text file https://www.matroska.org/technical/specs/subtitles/srt.html.

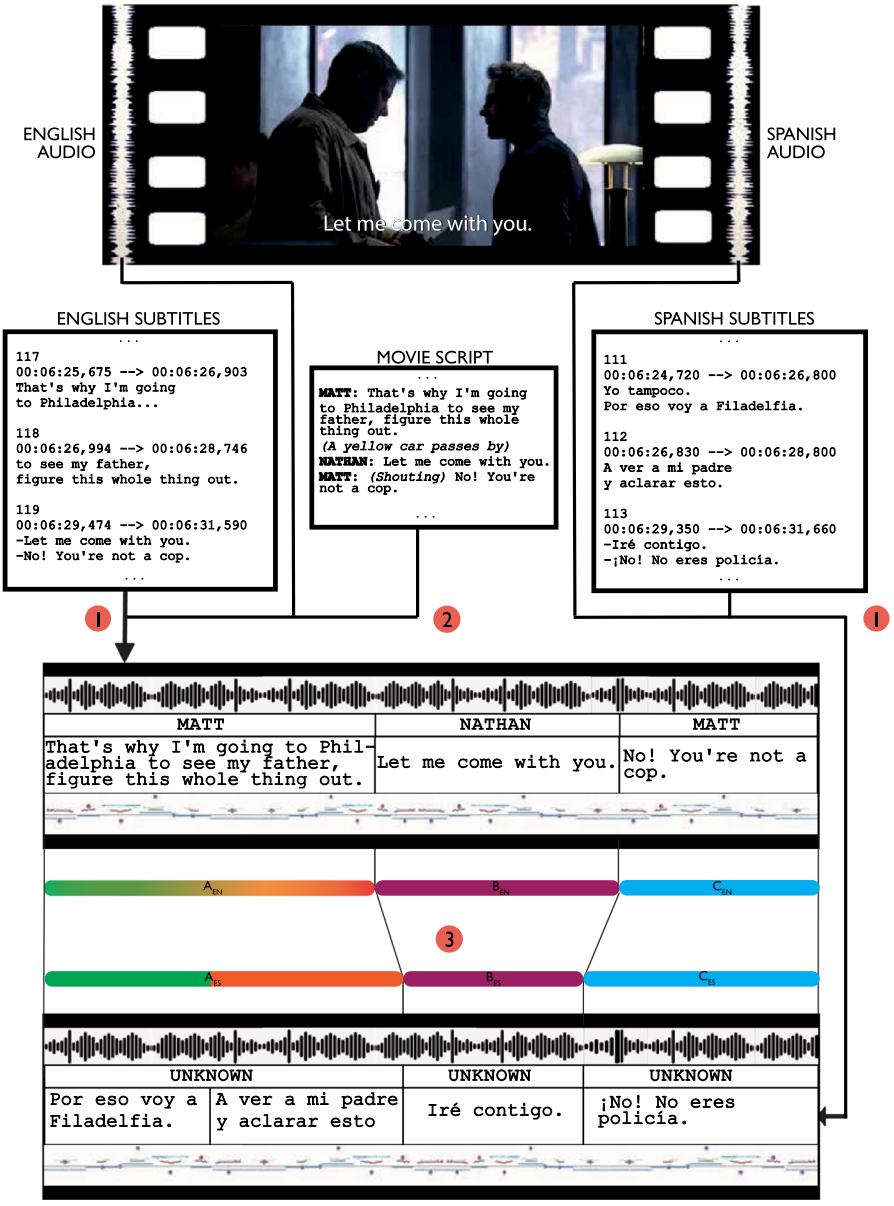

Fig. 5 Processes 1, 2 and 3 of the methodology illustrated on a portion of a movie: (1) Segment extraction using subtitle cues, (2) Speaker annotation from movie script, (3) Parallel segment extraction where aligning segments have matching colors. (Color figure online)

lowercase letter. Multi-speaker segments are split from the words following speech-dashes [–]. This process is marked with the label "1" on Fig. 5.

Movie scripts, which contain dialogue and scene information, are valuable pieces of information for determining the segment speaker labels. A heuristic approach based on word-matching is followed to obtain subtitle and script alignment. This process is marked with the label "2" on Fig. 5.

Afterwards, each word in the extracted segments is automatically annotated with the acoustic features needed for the purpose. For example in our case, we were especially interested in obtaining pauses between words for their use in machine translation (see Sect. 5.2).

In a later phase, the speech segments are aligned to create the parallel segment pairs. Since the subtitles and the number of extracted segments can differ between languages, the segment alignments can be one-to-one, one-to-many, many-to-one or many-to-many depending on the sentencing structure in the subtitles. To solve it, a metric is defined that measures the time correlation percentage between two sets of ordered segments and then maps them using a heuristic approach. This process is marked with the label "3" on Fig. 5.

3.3.2 Fair-use principles of copyrighted media

Movie and TV shows are protected with copyright laws and limit the amount of its usage. This is governed by the principles of *fair use*, which lets the use of copyrighted material for transformative and non-commercial purpose. The boundaries of what counts as transformative is not defined in a rigid way, but governed with guidelines and court decisions. The term "fair use" is originally defined by the United States law[13] and is influenced in other countries. United Kingdom, for example, allows non-commercial research on any material as long as it is within lawful access.[14]

Our methodology complies with these principals since the small audio portions obtained with it cannot be reconstructed back to the original form of the movie. The copyright on the original source of the segments has to be stated in both any publication explaining the work and during its access.

4 Building prosody corpora

This section deals with the description of methodologies used to build two different corpora for prosody-related applications. First, the processing of TED Talks to build a prosody-specific corpus; and second, the compilation of a movie-domain parallel corpus.

[13] https://www.copyright.gov/fair-use/more-info.html.

[14] https://www.gov.uk/guidance/exceptions-to-copyright.

Table 3 PANTED corpus statistics

Unit	Counts
# Talks	1046
# Speakers	884
# Hours recording	248:34:12
Avg. time/talk	0:14:15
# Sentences	156,407
# Words	2.4M
Avg. # words/sentence	15.06

4.1 Prosodically annotated TED talks (PANTED) corpus

TED (Technology, Entertainment, Design) talks are a set of conference talks lasting 15 min each in average, and held worldwide in more than 100 languages. TED talks, include a large variety of topics, from technology and design to science, culture and academia. The corresponding transcripts, as well as audio and video files, are available openly on TED's website.[15] For its public availability, TED talks have been the source of many corpora for linguistic analysis and machine learning-based applications. One example of this is the corpus-based study of paragraph-based prosodic cues by Farrús, Lai and Moore (2016). For their work, Farrús et al. compiled a corpus of 1365 talks together with their punctuated transcriptions and extracted various F0 and intensity based features at word and sentence levels.

Prosodically annotated TED talks (PANTED) corpus is the result of reprocessing of the corpus used in Farrús et al. (2016) to serve for joint processing of prosodic and other linguistic features in speech. Table 3 shows the main statistics of the corpus. PANTED corpus is made publicly available through the UPF e-repository[16] with Attribution 4.0 International (CC BY 4.0) license.[17] Source code used during the corpus reprocessing is also accessible online.[18]

4.2 The Heroes corpus

The *movie2parallelDB* methodology presented in the previous section has been put into practice by compiling a corpus from 2000s popular science fiction TV series *Heroes*.[19] Originating from United States, Heroes ran in TV channels worldwide between the years 2006 and 2010. The whole series consists of 4 seasons and 77 episodes and is dubbed into many languages including Spanish, Portuguese, French and Catalan. Each episode runs for a length of 42 min in average.

[15] http://www.ted.com.

[16] http://repositori.upf.edu/handle/10230/33981.

[17] https://creativecommons.org/licenses/by/4.0/.

[18] https://github.com/alpoktem/ted_preprocess.

[19] Produced by Tailwind Productions, NBC Universal Television Studio (2006–2007) and Universal Media Studios (2007–2010).

Twenty-one episodes from seasons 2 and 3 were processed using our methodology to obtain 7000 parallel audio segments together with their transcriptions and Proscript annotations. The total duration audio content is about 9.5 h. See Table 4 for further details. Counts of several linguistic units (words, tokens, sentences) in the final parallel corpus are presented in Table 5. A summary of how much of the content in one episode ended up in the dataset in average is presented in Table 6.

The dataset is published online as *Heroes Corpus*[20] and is accessible through Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license.[21] In what follows we further explain the additional processes that were followed during the collection of the dataset.

4.2.1 Raw data acquisition

The DVD's of the series were obtained from the Pompeu Fabra University Library. Videos of the episodes were extracted using the *Handbrake* video transcoder software[22] and were saved as *Matroska format (mkv)* files. To run *movie2parallelDB* scripts, we needed to extract audio and subtitle pairs for both languages residing in them. We used *MKVToolNix*[23] for extracting the audio. As subtitles were embedded as bitmap images in the DVD, we used an optical character recognition (OCR) software[24] to convert them to *srt* format subtitles. In total, 21 episodes were processed to obtain 25 h English and Spanish audio with their corresponding subtitles. The episode scripts were obtained from a fan web page.[25]

4.2.2 Manual subtitle correction work

We sourced the transcriptions of audio segments from the subtitles. Although subtitles are highly reliable sources for obtaining proper transcriptions in the original language of the movie, this is not always the case in the dubbed languages. This is due to the fact that dubbing transcript needs to satisfy visual alignment such as lip movements, whereas subtitles do not. Also, subtitles are often done in a more concise way to facilitate easy reading. In our case, we observed that the Spanish subtitles were matching with the Spanish audio in approximately 80% of the cases. As we needed exact correspondence between audio and transcriptions, we had to follow a manual correction process. Both subtitle transcripts and time-stamps had to be corrected to match exactly what is being spoken on the dubbing audio and when.

A manual passing over the segments also gave the opportunity to filter out any noisy audio portions that would otherwise end up in the corpus. Segments that contained noise and music, overlapping or unintelligible speech and speech in other

[20] http://hdl.handle.net/10230/35572.

[21] https://creativecommons.org/licenses/by-sa/4.0/.

[22] https://handbrake.fr/.

[23] https://mkvtoolnix.download/.

[24] Through a functionality provided by *Subler*: https://subler.org/.

[25] https://heroes-transcripts.blogspot.com/.

Table 4 Heroes corpus duration information

	English	Spanish
Total duration	4:45:36	4:43:20
Avg. duration/segment	00:02.44	00:02.42

Table 5 Word, token, sentence counts and average word count for parallel English and Spanish segments

Counts	English	Spanish
# Words	56,320	48,593
# Tokens	72,565	63,014
# Sentences	9892	9397
Avg. # words/sentence	5.69	5.17
Avg. # words/segment	8.04	6.94
Avg. # sentences/segment	1.41	1.34

Table 6 Average numbers for each episode

Counts	English	Spanish
Avg. # sentences (subtitles)	647	554
Avg. # sentences (extracted)	628	513
Avg. # segments	526	459
Avg. # parallel segments	334	

languages were marked. The spell checking and timestamps and script correction of 21 episodes was done by the two native Spanish speaking annotators and took 60 h in total.

5 Corpora evaluation

In this section we demonstrate the use of prosodic–lexical joint processing on two practical tasks using the corpora we have presented: automatic speech transcription and spoken language translation (Öktem, 2019). In the former case, we report experiments utilizing our data modeling methodology for the restoration of punctuation marks in raw automatic speech recognition output. In the latter case, we deal with enhancing spoken language translation through the incorporation of pause features.

5.1 Punctuation restoration using prosodic and lexical cues

The introduction of punctuation marks into the output of automatic speech recognition (ASR) is an important issue in applications such as automatic transcription/subtitling, speech-to-speech translation and language analysis.

Word sequence	he	who	knows	does	not	speak	he	who	speaks	does	not	know
Punctuation after	Ø	Ø	,	Ø	Ø	.	Ø	Ø	,	Ø	Ø	.

Fig. 6 modeling punctuation as a classification problem at each word interval (quote by *Lao Tze*)

Punctuation is essential for grammaticality, understandability, and functionality of several downstream tasks (Öktem et al., 2017a). For example, correct sentence segmentation and punctuation of recognized speech improves the quality of machine translation (Matusov, Mauser, & Ney, 2006; Peitz et al., 2011; Cho, Niehues, & Waibel, 2017; Lu & Ng, 2010), and missing periods and commas in machine generated text degrades performance of information extraction from speech (Favre et al., 2008; Hillard et al., 2006). Also, most of the data-driven parsing models require segmentation of recognized text into sentence-like units and use punctuation as features (Jones, 1994; Spitkovsky, Alshawi, & Jurafsky, 2011; Ma, Zhang, & Zhu, 2014).

During the manual transcription of an audio recording, both modalities, syntax and prosody, are used in determining the phrasing structure and punctuation. Analogously, an automatic system for restoring punctuation in automatic speech recognition output could take account of both of these modalities. Works that deal with the automatic punctuation restoration in ASR output demonstrated that prosodic features are highly indicative of phrase boundaries as well as of punctuation placement (Batista et al., 2012; Tilk & Alumäe, 2016; Xu, Xie, & Xiao, 2017).

In Öktem, Farrús and Wanner (2017a), we have proposed a framework that combines the processing of lexical and prosodic information for restoring punctuation in raw speech transcripts. The deep learning-based framework processes textual and prosodic information in a parallel fashion, reflecting the data structure we have in our datasets.[26]

The recurrent neural network (RNN)-based architecture treats the problem as a sequence prediction problem where a punctuation class is predicted at each interval of a sequence of words, as illustrated in Fig. 6. The proposed model makes it possible to integrate any desired feature (be it lexical, syntactic or prosodic) at each interval and allows us to test which feature influence punctuation placement accuracy to what extent. Our feature set includes word embeddings and part-of-speech (POS) tags as syntactic features, and fundamental frequency (F0), intensity, pauses and speech rate as prosodic features.

PANTED corpus was used as the training, development and testing dataset. Data was sampled into sequences of 50 words reflecting average number of words in the paragraph-segmentation structure in the corpus. Each sample starts with a new sentence. A total of 51,311 samples were extracted, consisting of 2.6 sentences in average. The train, validation and testing sets were split into percentages of 70–15–15 from the complete set.

[26] Code repository available at: https://github.com/alpoktem/punkProse.

Table 7 Punctuation generation results for each punctuation mark as F_1 score in percentage (%). Bold indicates highest results

Feature set	Comma (,)	Full stop (.)	Question (?)	Overall
Baseline: word (w)	53.4	77.6	65.9	65.9
w + pause (p)	52.2	79.3	67.4	67.2
$w + pos$	53.0	79.3	67.1	67.4
$w + pos + p$	53.7	80.5	**71.8**	68.2
$w + pos + p + f0$	**55.2**	81.9	69.7	**69.2**
$w + pos + p + i$	54.3	81.1	69.0	68.6
$w + pos + p + sr$	51.9	80.3	67.8	67.9
$w + pos + p + f0 + i$	51.9	**82.0**	69.8	68.8
$w + pos + p + f0 + sr$	54.0	81.5	68.1	68.6
$w + pos + p + f0$ (discrete)	54.4	**83.0**	71.2	**70.3**

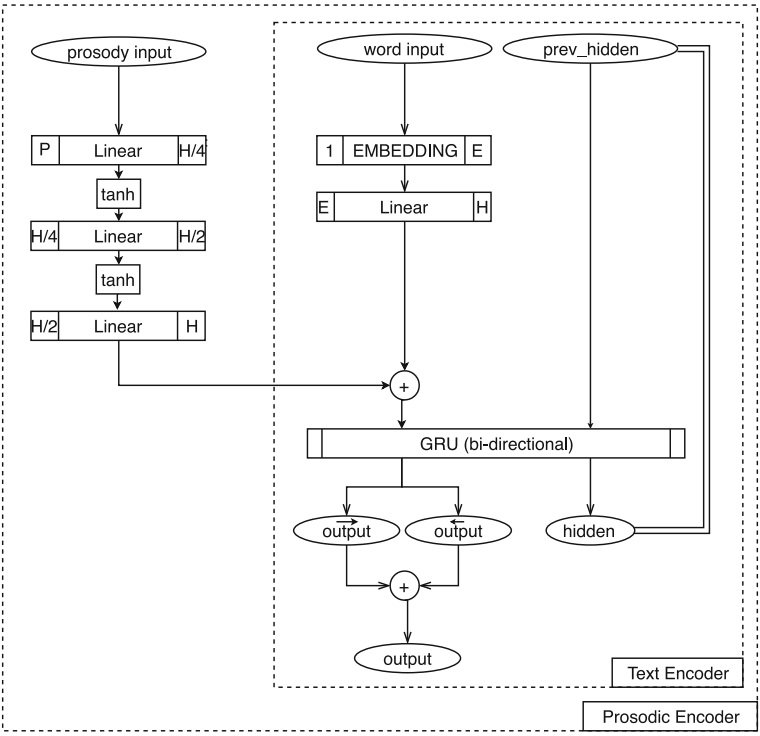

Fig. 7 Sequence-to-sequence translation encoder with prosodic information

Table 7 shows the results in generating full stops, commas and question marks with different feature settings with the TED test set. Features are denoted with symbols *p* (pause duration), *f*0 (intonation), *i* (intensity) and *sr* (speech rate). All reported results except the last row use prosodic features as continuous values in order to ensure comparability with the baseline architecture, replicated from Tilk and Alumäe (2016) and using only lexical information (*w*). For the final setup, we determined discrete levels for the *f*0 and *i* features and used them instead of their real values. This resulted in an improvement in terms of F_1 scores achieving 70.3% accuracy averaged over all three punctuation marks.

We can observe that all combination sets outperform the results obtained with the baseline architecture. The use of words, POS, pause duration and mean *F*0 features especially serves for detecting commas and all three punctuation marks overall. The same combination of features plus mean intensity values improve especially in the full stop detection. Only words, POS and pauses give the best *F*1-score when marking sentence boundaries that signal a question.

5.2 Enhancing spoken language translation with prosody

In this section, we explore the introduction of prosodic features directly on a deep learning-based machine translation system. We particularly focus on the inclusion of inter-lexical silent pauses as an additional feature on the encoder side of the sequence-to-sequence MT architecture (Sutskever, Vinyals, & Le, 2014; Öktem, 2019).

The enhanced encoder architecture is illustrated in Fig. 7. The text encoder (inner box) encodes the words by passing them as embedding vectors to a bidirectional recurrent layer. Meanwhile, a separate encoding sequence is followed for the pause features where they are gradually converted to be fit as input to the same recurrent layer concatenated with the word vector. Output vectors at each timestep are then passed on through an attentional decoder (Bahdanau, Cho, & Bengio, 2014), which only outputs words.[27]

Training is performed in two stages with two types of data: (1) a parallel text corpus and (2) a prosodically annotated parallel spoken audio corpus. In the first stage, training is performed updating only the parameters belonging to the text encoder and decoder. On a second stage, training is performed with the joint text and prosody encoder components. For base parallel text data, we used 5 million English–Spanish sentence pairs from the *OpenSubtitles corpus* (Lison & Tiedemann, 2016; Lison et al., 2018). As for the second stage training, we used a version of *Heroes corpus* consisting of 7225 parallel segments. This set was split into training–test–validation sets of size 6141, 542, 541 segments, respectively.

Table 8 shows the difference in translation quality metrics between the baseline (*text*) model and the model enhanced with pause information (*text* + *pauses*) on the two versions of test sets from Heroes corpus, one including the original punctuation and another including automatically restored punctuation simulating an automatic transcription scenario. In order to compare the different systems, we used BLEU

[27] Code repository available at: https://github.com/alpoktem/TransProse.

Table 8 BLEU scores (%) on the Heroes corpus testing set with and without pause encoding

Punctuation in input	Translation encoder type	
	Text	Text + pauses
Subtitle	20.15	21.46
Recovered	18.08	19.15

(Papineni et al., 2002), which is a widely-accepted automatic MT evaluation metric that is known to correlate with human judgements. With manually annotated punctuation on the input sentences, there is an improvement of 1.31% in terms of BLEU scoring. With punctuation recovery preprocessing on the raw transcripts, translation quality still increases by 1.07%. These improvements indicate a possibility of increase in neural machine translation quality by means of joint lexical–prosodic encoding.

6 Conclusion

The motivation to enhance speech processing applications with prosodic heuristic comes with a cost and that mostly resides in the labour of data harvesting. In this paper, we addressed the need for a tool set to obtain speech data with prosody heuristic and also presented the two corpora resulting from this motivation. We introduced a complete pipeline for collecting, handling, storage and visualization of this type of data. *Proscript* library served for handling of linguistically-aligned prosodic data. *Prosograph* enabled manual examination of this type of data by visualizing speech-related characteristics through a programmable interface. It can be used in many areas of research such as language learning and acquisition, comparative studies in different languages, tone languages, and audiovisual prosody, among others. Finally, the *movie2parallelDB* framework was built to create structured parallel speech data from dubbed movies.

The two data resources prepared and packaged using these tools are also presented, and include: (1) PANTED, consisting of automatically generated lexical–prosodic annotations of TED conference talks, and (2) the Heroes corpus, consisting of prosodically annotated parallel TV–movie domain speech segments. Using the latter one, we were able to demonstrate the increase in the accuracy of automatic punctuation restoration through the use of prosody heuristic. The former resource, to our knowledge, is the first example of a parallel movie speech corpus paving the way for research in dubbing translation (Öktem et al., 2019; Nayak et al., 2020). We furthermore outlined a deep learning-based machine translation framework where such corpora would be useful in integrating prosodic features into the neural machine translation logic.

All the developed resources, corpora, as well as evaluation frameworks are published openly. We hope that both the toolkit and corpora serves the speech technology and prosody research community.

Acknowledgements This work was largely done during the doctoral studies of the first author in Universitat Pompeu Fabra. The second author has been funded by the Agencia Estatal de Investigación (AEI), Ministerio de Ciencia, Innovación y Universidades and the Fondo Social Europeo (FSE) under Grant RYC-2015-17239 (AEI/FSE, UE).

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

References

Adami, A. G., Mihaescu, R., Reynolds, D. A., & Godfrey, J. J. (2003). Modeling prosodic dynamics for speaker recognition. In *2003 IEEE international conference on acoustics, speech, and signal processing, 2003. Proceedings (ICASSP'03)* (Vol. 4, pp. IV-788). IEEE.

Almeman, K., Lee, M., & Almiman, A.A. (2013). Multi dialect Arabic speech parallel corpora. In *1st International conference on communications, signal processing, and their applications (ICCSPA)* (pp. 1–6). IEEE.

Avanzi, M., Lacheret-Dujour, A., & Victorri, B. (2008). ANALOR. A tool for semi-automatic annotation of French prosodic structure.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. CoRR.

Batista, F., Moniz, H., Trancoso, I., & Mamede, N. (2012). Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(2), 474–485.

Bendazzoli, C., & Sandrelli, A. (2005). An approach to corpus-based interpreting studies: Developing EPIC (European Parliament Interpreting Corpus). In *MuTra 2005–Challenges of multidimensional translation* (pp. 1–12).

Boersma, P., & Weenink, D. (2019). Praat: Doing phonetics by computer (computer program), version 6.0.46. Retrieved January 3, 2019, from http://www.praat.org/ .

Canavan, A., Graff, D., & Zipperlen, G. (1997). *CALLHOME American English Speech LDC97S42*. Web download. Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC97S42 .

Cettolo, M., Girardi, C., & Federico, M. (2012). Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the international conference of the European Association for Machine Translation (EAMT)*, Trento, Italy (pp. 261–268).

Cho, E., Niehues, J., & Waibel, A. H. (2017). NMT-based segmentation and punctuation insertion for real-time spoken language translation. In *Proceedings of Interspeech 2017*.

Crystal, D. (2003). *A dictionary of linguistics and phonetics*. Blackwell Publishing Ltd.

Domínguez, M., Farrús, M., & Wanner, L. (2016a). An automatic prosody tagger for spontaneous speech. In *Proceedings of the 26th international conference on computational linguistics (COLING)*, Osaka, Japan (pp. 377–386).

Domínguez, M., Latorre, I., Farrús, M., Codina-Filbà, J., & Wanner, L. (2016b). Praat on the web: An upgrade of Praat for semi-automatic speech annotation. In *Proceedings of the 26th international conference on computational linguistics (COLING)*, Osaka, Japan (pp. 218–222).

Farrús, M., Lai, C., & Moore, J. D. (2016). Paragraph-based cues for speech synthesis applications. In *Proceedings of the speech prosody*, Boston, MA.

Favre, B., Grishman, R., Hillard, D., Ji, H., Hakkani-Tur, D., & Ostendorf, M. (2008). Punctuating speech for information extraction. In *IEEE international conference on acoustics, speech and signal processing, 2008. ICASSP 2008* (pp. 5013–5016). IEEE.

Federmann, C., & Lewis, W. D. (2016). Microsoft Speech Language Translation (MSLT) corpus: The IWSLT 2016 release for English, French and German. In *International workshop on spoken language translation*.

Fujisaki, H. (1997). Prosody, models, and spontaneous speech. In Y. Sagisaka, N. Campbell & N. Higuchi (Eds.), *Computing prosody: Computational models for processing spontaneous speech* (pp. 27–42). Springer.

Garrido, J., Codina, M., & Fodge, K. (2018). TransDic, a public domain tool for the generation of phonetic dictionaries in standard and dialectal Spanish and Catalan. In *Proceedings of Iberspeech*, Barcelona, Spain (pp. 291–295).

Georgescu, A. L., Cucu, H., Buzo, A., & Burileanu, C. (2020). RSC: A Romanian read speech corpus for automatic speech recognition. In *Proceedings of the 12th language resources and evaluation conference* (pp. 6606–6612).

Godfrey, J., & Holliman, E. (1993). *Switchboard-1 release 2 ldc97s62*. DVD. Linguistic Data Consortium.

Hermann, K. M., & Blunsom, P. (2014). Multilingual models for compositional distributional semantics. In *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (ACL)*, Baltimore, Maryland, USA (pp. 58–68).

Hillard, D., Huang, Z., Ji, H., Grishman, R., Hakkani-Tur, D., Harper, M., Ostendorf, M., & Wang, W. (2006). Impact of automatic comma prediction on POS/name tagging of speech. In *Proceedings of the IEEE spoken language technology workshop*, Palm Beach, Aruba (pp. 58–61).

Huang, Z., Chen, L., & Harper, M. (2006). An open source prosodic feature extraction tool. In *Proceedings of the fifth international conference on language resources and evaluation (LREC)*, Genoa, Italy.

Jones, B. E. M. (1994). Exploring the role of punctuation in parsing natural text. In *Proceedings of the 15th conference on computational linguistics, COLING '94* (Vol. 1, pp. 421–425). Association for Computational Linguistics. https://doi.org/10.3115/991886.991960.

Külebi, B., & Öktem, A. (2018). Building an open source automatic speech recognition system for Catalan. In *Proceedings of IberSPEECH 2018* (pp. 25–29). https://doi.org/10.21437/IberSPEECH.2018-6.

Külebi, B., Öktem, A., Peiró-Lilja, A., Pascual, S., & Farrús, M. (2020). CATOTRON—A neural text-to-speech system in Catalan. In *Proceedings of Interspeech 2020* (pp. 490–491).

Lacheret, A., Kahane, S., Beliao, J., Dister, A., Gerdes, K., Goldman, J., Obin, N., Pietrandrea, P., & Tchobanov, A. (2014). Rhapsodie: A prosodic-syntactic treebank for spoken French. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (Eds.), *Proceedings of the ninth international conference on language resources and evaluation, LREC 2014*, Reykjavik, Iceland, May 26–31, 2014 (pp. 295–301). European Language Resources Association (ELRA).

Lison, P., & Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *LREC 2016, tenth international conference on language resources and evaluation*. European Language Resources Association.

Lison, P., Tiedemann, J., & Kouylekov, M. (2018). Open subtitles 2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *LREC 2018, eleventh international conference on language resources and evaluation*. European Language Resources Association (ELRA).

Lu, W., & Ng, H. T. (2010). Better punctuation prediction with dynamic conditional random fields. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 177–186). Association for Computational Linguistics.

Ma, J., Zhang, Y., & Zhu, J. (2014). Punctuation processing for projective dependency parsing. In *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics: Short papers* (Vol. 2, pp. 791–796). Association for Computational Linguistics. http://www.aclweb.org/anthology/P14-2128.

Matusov, E., Mauser, A., & Ney, H. (2006). Automatic sentence segmentation and punctuation prediction for spoken language translation. In *International workshop on spoken language translation (IWSLT) 2006*.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2013). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of tools and resources for the analysis of speech prosody (TRASP)*, Aix-en-Provence, France (pp. 7–10).

Mertens, P. (2004). The Prosogram: Semi-automatic transcription of prosody based on a tonal perception model. In *Proceedings of the 2nd international conference on speech prosody*, Nara, Japan (pp. 549–552).

Nayak, S., Baumann, T., Bhattacharya, S., Karakanta, A., Negri, M., & Turchi, M. (2020). See me speaking? Differentiating on whether words are spoken on screen or off to optimize machine dubbing. In K. P. Truong, D. Heylen, M. Czerwinski, N. Berthouze, M. Chetouani & M. Nakano (Eds.), *Companion publication of the 2020 international conference on multimodal interaction, ICMI companion 2020, virtual event*, The Netherlands, October, 2020 (pp. 130–134). ACM. https://doi.org/10.1145/3395035.3425640.

Öktem, A. (2019). Incorporating prosody into neural speech processing pipelines: Applications on automatic speech transcription and spoken language machine translation. PhD Thesis, Universitat Pompeu Fabra.

Öktem, A., Farrús, M., & Bonafonte, A. (2018). Bilingual prosodic dataset compilation for spoken language translation. In *Proceedings of the Iberspeech*, Barcelona, Spain (pp. 20–24).

Öktem, A., Farrús, M., & Bonafonte, A. (2019). Prosodic phrase alignment for machine dubbing. In G. Kubin & Z. Kacic (Eds.), *Interspeech 2019, 20th annual conference of the International Speech Communication Association*, Graz, Austria, September 15–19, 2019 (pp. 4215–4219). ISCA. https://doi.org/10.21437/Interspeech.2019-1621.

Öktem, A., Farrús, M., & Wanner, L. (2017a). Attentional parallel RNNs for generating punctuation in transcribed speech. In *Statistical language and speech processing* (pp. 131–142). Springer.

Öktem, A., Farrús, M., & Wanner, L. (2017b). Prosograph: A tool for prosody visualisation of large speech corpora. In *Proceedings of Interspeech*, Stockholm, Sweden (pp. 809–810).

Ostendorf, M., Price, P., & Shattuck-Hufnagel, S. (1996). *Boston University radio speech corpus LDC96S36*. DVD. Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC96S36.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on Association for Computational Linguistics (ACL)*, Philadelphia, Pennsylvania (pp. 311–318).

Pappas, N., & Popescu-Belis, A. (2013) Sentiment analysis of user comments for one-class collaborative filtering over TED talks. In *Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval*, Dublin, Ireland (pp. 773–776).

Parada-Cabaleiro, E., Costantini, G., Batliner, A., Schmitt, M., & Schuller, B. W. (2019). DEMoS: An Italian emotional speech corpus. *Language Resources and Evaluation*, *54*, 1–43.

Peitz, S., Freitag, M., Mauser, A., & Ney, H. (2011). Modeling punctuation prediction as machine translation. In *International workshop on spoken language translation (IWSLT) 2011*.

Rosenberg, A. (2010). AuToBi—A tool for automatic ToBi annotation. In *Eleventh annual conference of the International Speech Communication Association*.

Rousseau, A., Deléglise, P., & Estève, Y. (2012). TED-LIUM: An automatic speech recognition dedicated corpus. In *Proceedings of the eighth international conference on language resources and evaluation (LREC)*, Istanbul, Turkey.

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category—ELAN and ISO DCR. In *6th International conference on language resources and evaluation (LREC 2008)*.

Spitkovsky, V. I., Alshawi, H., & Jurafsky, D. (2011). Punctuation: Making a point in unsupervised dependency parsing. In *Proceedings of the fifteenth conference on computational natural language learning, CoNLL '11* (pp. 19–28). Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2018936.2018939.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th international conference on neural information processing systems—NIPS'14* (Vol. 2, pp. 3104–3112). MIT Press.

Takamichi, S., & Saruwatari, H. (2018). CPJD corpus: Crowdsourced parallel speech corpus of Japanese dialects. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L18-1067.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *LREC* (Vol. 2012, pp. 2214–2218).

Tilk, O., & Alumäe, T. (2016). Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *Proceedings of Interspeech*, San Francisco, CA, USA (pp. 3047–3051).

Tsiartas, A., Ghosh, P., Georgiou, P. G., & Narayanan, S. (2011). Bilingual audio-subtitle extraction using automatic segmentation of movie audio. In *Proceedings of the international conference on acoustics, speech and signal processing (ICASSP)*, Prague, Czech Republic (pp. 5624–5627).

Wester, M. (2010). *The EMIME Bilingual Database*. Technical report. The University of Edinburgh.

Xu, Y. (2013). ProsodyPro—A tool for large-scale systematic prosody analysis. In *Proceedings of tools and resources for the analysis of speech prosody (TRASP)*, Aix-en-Provence, France (pp. 7–10).

Xu, C., Xie, L., & Xiao, X. (2017). A bidirectional LSTM approach with word embeddings for sentence boundary detection. *Journal of Signal Processing Systems*, *90*(7), 1063–1075. https://doi.org/10.1007/s11265-017-1289-8.

Zanon Boito, M., Havard, W. N., Garnerin, M., Le Ferrand, É., & Besacier, L. (2020). MaSS: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the Bible. In *Proceedings of the 12th language resources and evaluation conference*, Marseille, France (pp. 6486–6493). https://hal.archives-ouvertes.fr/hal-02611059.