Automatic Extraction of Parallel Speech Corpora from Dubbed Movies

Alp Öktem, Mireia Farrús, Leo Wanner
Presented by: Simon Mille

{alp.oktem|mireia.farrus|leo.wanner|simon.mille}@upf.edu
1. Parallel Speech Corpora
2. Movies as a resource for parallel corpora
3. Proposed Method
4. Methodology
5. Applying the methodology
6. Discussions
7. Conclusions
8. Future work
Spoken parallel corpora are useful in building speech-to-speech applications.

Costly: Laborious with respect to translation and interpretation

Available corpora:

- Contain unexpressive speech (e.g. interpreted)
- Do not capture spontaneous spoken language traits (e.g. read)
- Lack one-to-one alignment between words/sentences (e.g. constrained conversations)
Dubbed movies as a resource

Popular movies, documentaries, TV shows are dubbed in many countries*.

A good resource for obtaining bilingual data:

(1) Available parallel audio data in dubbed movies.
(2) Transcripts available with time information in subtitles.

| Time | Subtitle (ENG)                                   |
|------|--------------------------------------------------|
| Ss1  | Se1 ...table. (LAUGHING)                        |
| Ss2  | Se2 (HORN HONKING)                              |
| Ss3  | Se3 Hey, look! A camel is sitting               |
| Ss4  | Se4 there.                                      |

| Time | Subtitle (SPA)                                   |
|------|--------------------------------------------------|
| Ss1  | Se1 ...mesa. (RIENDO)                            |
| Ss2  | Se2 (BOCINA)                                     |
| Ss3  | Se3 Miren! Un camello                            |
| Ss4  | Se4 esta sentado allí.                           |

Movie still from *The Man Who Knew Too Much* (1956) © Universal Pictures
Example: Dubbing in European Countries

European countries and their common methods to dub films

- Dubbing only for children: Otherwise solely subtitles
- Mixed areas: Countries using occasionally full-cast dubbing otherwise solely subtitles
- Voice-over: Countries using usually one or just a couple of voice actors whereas the original soundtrack persists
- General dubbing: Countries using exclusively a full-cast dubbing, both for films and for TV series
- Countries which occasionally produce own dubbings but generally using dubbing versions of other countries since their languages are quite similar to each other and the audience is also able to understand it without any problems. (Belgium and Slovakia)

Image source: Wikipedia - Dubbing (filmmaking)
Proposed Method

Automatic extraction of segmented parallel sentences with prosodic parameters

➔ Input: Bilingual audio and subtitles pair
➔ Output: Aligned bilingual sentences annotated with prosodic features

Key points:

1. Supports any language pair
2. Contains expressive speech
3. Aligned at sentence level
Methodology Overview

Monolingual corpus creation in parallel

(1) Sentence Segmentation

(2) Prosodic Parameter Extraction

(3) Parallel corpus creation

\( \text{lang}_A \) audio\(_A\) subtitle\(_A\)

\( \text{lang}_B \) audio\(_B\) subtitle\(_B\)
Stage 1: Sentence Segmentation

Use subtitle time-information to find script location in audio

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.
Stage 1: Sentence Segmentation

subtitle cleaning + crude cutting

audio/transcript alignment

sentence detection

parallel sentence mapping

| pair id | English       | Spanish        |
|---------|---------------|----------------|
| n       | Hey, look!    | Miren!         |
| n + 1   | A camel is sitting there. | Un camello esta sentando allí. |

1VoxSigma® Scribe.
Stage 2: Prosodic Parameter Extraction

ProsodyPro\(^1\) library used for prosodic feature extraction

\(^1\)Yi Xu. 2013
Stage 3: Parallel Sentence Alignment

Goal: Given sentence $s_1$ in lang. 1 find corresponding sentence $s_2$ in lang.2

![Diagram of parallel sentence alignment]

| Pair ID | SPA                          | ENG                          |
|---------|------------------------------|------------------------------|
| 1       | Yo soy Ambrose Chappell      | I am Ambrose Chappell.       |
| 2       | ¿Qué se le ofrece?           | What can I do for you?      |
| 3       | Podríamos empezar con su nombre. | If you gave me your name, that might be a start. |
| 4       | Me llamo McKenna.            | My name is McKenna.          |
| 5       | Soy el Doctor Benjamín McKenna. | Doctor Benjamin McKenna.     |

1. Yandex Translate
2. Meteor library (Denkowski and Lavie, 2014)
Applying the Methodology

Three movies processed:

- The Man Who Knew Too Much (1956)
- Slow West (2015)
- The Perfect Guy (2015)

Films originally in English, dubbed to Spanish.

Audio extracted from DVD using Libav\(^1\)

English and Spanish subtitles obtained from opensubtitles\(^2\).

\(^1\)https://libav.org/
\(^2\)https://www.opensubtitles.org
### Currently obtained corpus

| 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                          |
Shortcomings

➢ Copyright restrictions for distributing the corpus.

Main bottlenecks in capturing data:

1. Audio-text alignment performance
   ○ 15% sentences lost in original language.
   ○ 49% sentences lost in dubbed language.

2. Translation difference in dubbed audio and subtitles
   ○ Hinders audio-text alignment

3. Background noise

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

Processing *The Man Who Knew Too Much*
| Extracted | English (Sub + audio) | Spanish Sub | Spanish Dub |
|-----------|-----------------------|-------------|-------------|
| yes       | Daddy, you’re sure I’ve never been to Africa before? | Papá, ¿estás seguro de que nunca estuve antes en África? | Papa, estás seguro que no habíamos estado ya en África? |
| no        | It looks familiar. | Me parece conocido. | Todo esto ya lo conozco. |
| yes       | You saw the same scenery last summer driving to Las Vegas. | Viste el mismo panorama el verano pasado cuando manejamos a Las Vegas. | Vimos un paisaje muy parecido cuando fuimos a Las Vegas. |
| yes       | Where Daddy lost all that money at the crap ... | Claro, donde papá perdió todo ese dinero en la mesa | Ah, claro, donde papá perdió toda el dinero en la mesa de juego? |
| no        | Hey, look! | ¡Miren! | Hey mirad! |
| no        | A Camel. | ¡Un camello! | Un camello! |
| yes       | Of course this isn't really Africa, honey. | Y esto no es realmente África. | Realmente esto no es África, cariño. |
| yes       | It's the French Morocco. | Es el Marruecos francés. | Es el Marruecos Francés. |
| yes       | Well, it's northern Africa. | Es África del Norte. | Bueno, es África del Norte. |
| yes       | Still seems like Las Vegas. | Aún se parece a Las Vegas. | Pues, sigue pareciéndose a Las Vegas. |
Conclusions

- Automatic building of multimodal bilingual corpora from dubbed media
  - Speech, text, prosody
  - Conversational speech → Useful for speech-to-speech translation applications
- Works on any language pair (with trained acoustic model)
- No further training needed
- Code available at http://www.github.com/TalnUPF/movie2parallelDB
Future Work

1. Switch from proprietary audio-text aligner software to open source
   ➢ E.g. p2fa (based on CMU Sphinx ASR system)

2. XML based structure as corpus metadata
   ➢ Instead of directory structure only

3. Speaker diarization
   ➢ Identifying the speaker of each sentence

4. Extend and publish the corpus
   ➢ Depending on agreement with Copyright holders
## Appendix A: State of the Art Corpora

| 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.
# Appendix B: ProsodyPro Files

| ProsodyPro output file | Description |
|------------------------|-------------|
| rawf0                  | Raw $f_0$ contour in Hz |
| f0                     | Smoothed $f_0$ with trimming algorithm (Hz) |
| smoothf0               | Smoothed $f_0$ with triangular window (Hz) |
| semitonef0             | $f_0$ contour in semitones |
| samplef0               | $f_0$ values at fixed time intervals (Hz) |
| f0velocity             | First derivative of $f_0$ |
| means                  | $f_0$, intensity and velocity parameters (mean, max, min) for each word |
| normtimef0             | Constant number of $f_0$ values for each word |
| normtimeIntensity      | Constant number of intensity values for each word |

Some of the files generated by ProsodyPro