A first summarization system of a video in a target language
The key challenge and potential impact

- How to make the main idea presented in a video in a foreign language accessible and easy to understand by everyone?
- Accessing to information in foreign languages would permit to access to the other side of a story.
- Due to political, socio-cultural or religion reasons, divergence of opinions may exist within two medias from two different sources.
Objective 1: Understanding the main idea of a video by summarizing

- **Input**: A video in Arabic or French.
- **Output**: A summary of the input video subtitled in English. This summary is supposed to capture the main idea of the source video.
Objective 2: Cross-lingual opinion Analysis

- Input: An Arabic video will be compared to French or English video.
- Output: A review concerning the degree of divergence between the two videos.
Global architectures

To reach the first objective several components are necessary. We developed several scenarios.
What is necessary to develop AMIS?

1. Collecting video data
2. Video Summarization
3. Speech recognition system
4. Machine Translation
5. Text and Audio summarization
6. Making multilingual corpora comparable
7. Labeling (sentiments and opinions) the multilingual corpora
8. Evaluation (objective and subjective)
1- Video database

Data on conflicting topics for studying the cross-lingual opinions in Arabic, English and French.

<table>
<thead>
<tr>
<th>Syria</th>
<th>Real Madrid - FC Barcelona</th>
<th>Animal rights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women’s rights</td>
<td>Homosexual marriage</td>
<td>Drug liberalization</td>
</tr>
<tr>
<td>Death sentence</td>
<td>Occupied territories</td>
<td>Trump</td>
</tr>
</tbody>
</table>

**Arabic channels**

**English channels**

**French channels**
Some statistics

- 100 hours of videos have been collected for each of the three languages.
- For each video in a language, videos in the 2 other languages correspond to the same topic.
2- Video Summarization

- Anchor-person.
- Recognition of day and night shots.
- These indicators are used for calculating coefficient for activity.
- The coefficient of activity is calculated for each frame and then as an average per each shot.
Jump Cuts and Dissolve (Fade) to White

Jumpcut – cut in film editing in which two sequential shots of the subject taken from camera positions varying only slightly if at all

Dissolve (fade) to white used to soften jump cuts startling viewer
Arabic is considered as the foreign language. Development of an Arabic automatic speech recognition system with the following data:

**Acoustic corpus (hours):** 63 hours.

- 35 acoustic models have been trained
- The emission probabilities of the HMM models are estimated by DNN.
- The number of parameters to estimate is about 30 million.

**Text corpus**

- GigaWord corpus was collected from 9 sources of information with a total of $10^9$ words.
- 315K words from the transcription of the acoustic training data
- A vocabulary of 95K words with an average of 5.07 pronunciations for each entry
A statistical Machine Translation has been developed Arabic-English.

For the training process, we use a parallel corpus extracted from UN proceedings of 9.7 million of sentences.

For the target language we used a 4-gram language model

The vocabulary contains 224000 entries.
Several techniques of text summarization do exist in the literature.

- Summarization by extraction
- Summarization by abstraction
- Summarization by sentence compression

Our purpose is to summarize the result of a noisy text. We decided mainly to adopt the extractive summarization paradigm in AMIS project.
5- Audio Summarization

- Selecting segments within the audio signal that are considered more relevant.
- Informativeness of each segment is obtained by mapping a set of audio features issued from its Mel-frequency Cepstral Coefficients and their corresponding Jensen-Shannon divergence score.
First we need to make all the database comparable.

- To compare two videos in two different languages in terms of sentiment, we need first to make them comparable.
- We tested several based-dictionary methods for the comparability.
- Each document is represented by a TFIDF vector.
- The dimension $n$ of a vector is fixed in terms of the most frequent words in Arabic.
- An English document is represented also by a vector of dimension $n$ corresponding to the same words such as in Arabic.
Making corpora comparable

- The evaluation is done on 123 videos of Euronews that we know that they are comparable.

<table>
<thead>
<tr>
<th></th>
<th>Max</th>
<th>Min</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>359</td>
<td>39</td>
<td>160</td>
</tr>
<tr>
<td>English</td>
<td>432</td>
<td>38</td>
<td>186</td>
</tr>
</tbody>
</table>

**Table**: Statistics in terms of words of the test corpus

- The method achieved a score of a Recall of 0.73 and a Precision of 0.62
We asked 12 annotators to label videos in terms of sentiments (polarity).

They label 40 videos.

Each transcribed document has been segmented into phrases.

The used labels are the following:

- 1: If the phrase is Positive
- -1: If the phrase is Negative
- 0: If the sentence is Neutral
- ∞: No Understandable
Results of Labeling in terms of videos (40)

(a) Arabic
- Neutral: 47.5%
- Positive: 10%
- Negative: 40%
- No Understandable: 2.5%

(b) English
- Neutral: 42.5%
- Positive: 12.5%
- Negative: 42.5%
- No Understandable: 2.5%
Evaluation of the different components

In the following table, each component of the system has been evaluated.

<table>
<thead>
<tr>
<th>Compo</th>
<th>R</th>
<th>P</th>
<th>WER</th>
<th>BLEU</th>
<th>Test Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video Sum</td>
<td>0.13</td>
<td>0.36</td>
<td>X</td>
<td>X</td>
<td>50 Seq</td>
</tr>
<tr>
<td>ALASR</td>
<td>X</td>
<td>X</td>
<td>14.02</td>
<td>X</td>
<td>31K Sent</td>
</tr>
<tr>
<td>MT</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>0.39</td>
<td>3K Sent</td>
</tr>
</tbody>
</table>

These results have been achieved on data not extracted from our video database except for the video summarization.
Evaluation on AMIS data

- **ASR**: Since there is no reference, we considered the transcription of Youtube and Euonews as reference. On a corpus test of 1300 sentences we get a WER of 36.5.

- **MT**: There is no reference, we translated the output of the ASR with Google and we considered it as a reference. The value of BLEU on a test set of 197 videos is 26.7.
End User Evaluation of Improved Components for Speech Recognition, Machine Translation and Video/Audio/Text Summarisation

Scenario 1 - 45%
Scenario 3 - 26%
Scenario 4 - 29%

Figure: Subjective evaluation losers among 11 professionals
Evaluation of the quality of an Audio summary

Five people were asked to evaluate the quality of a summary as a whole and the quality of each segment independently from the whole summary.

<table>
<thead>
<tr>
<th>Score</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Full informative</td>
</tr>
<tr>
<td>4</td>
<td>Mostly informative</td>
</tr>
<tr>
<td>3</td>
<td>Half informative</td>
</tr>
<tr>
<td>2</td>
<td>Quite informative</td>
</tr>
<tr>
<td>1</td>
<td>Not informative</td>
</tr>
</tbody>
</table>

Table: Evaluation scale
## Results

<table>
<thead>
<tr>
<th>Sample</th>
<th>Length</th>
<th>Segments</th>
<th>Full Score</th>
<th>Average Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3m19s</td>
<td>8</td>
<td>4.20</td>
<td>2.90</td>
</tr>
<tr>
<td>2</td>
<td>5m21s</td>
<td>13</td>
<td>3.50</td>
<td>2.78</td>
</tr>
<tr>
<td>3</td>
<td>2m47s</td>
<td>5</td>
<td>3.80</td>
<td>3.76</td>
</tr>
<tr>
<td>4</td>
<td>1m42s</td>
<td>5</td>
<td>3.60</td>
<td>2.95</td>
</tr>
<tr>
<td>5</td>
<td>8m47s</td>
<td>22</td>
<td>4.67</td>
<td>3.68</td>
</tr>
<tr>
<td>6</td>
<td>9m45s</td>
<td>30</td>
<td>4.00</td>
<td>2.49</td>
</tr>
<tr>
<td>7</td>
<td>5m23s</td>
<td>8</td>
<td>3.20</td>
<td>3.75</td>
</tr>
<tr>
<td>8</td>
<td>6m24s</td>
<td>20</td>
<td>3.75</td>
<td>2.84</td>
</tr>
<tr>
<td>9</td>
<td>7m35s</td>
<td>18</td>
<td>3.75</td>
<td>3.19</td>
</tr>
<tr>
<td>10</td>
<td>2m01s</td>
<td>4</td>
<td>2.75</td>
<td>2.63</td>
</tr>
</tbody>
</table>

**Table:** Audio summarization performance over complete summaries and summary segments
The initial idea becomes a reality and we developed a prototype.

New scientific challenges arise: Audio summarization, Code-switched signal (speech recognition and machine translation), several people speaking simultaneously, etc.

Subjective evaluation is under progress.

Deployment of other architectures.

Comparison of opinions.