Extractive Text-Based Summarization of Arabic videos: Issues, Approaches and Evaluations

M.A. Menacer, C.E. González-Gallardo K. Abidi D. Fohr D. Jouvet D. Langlois O. Mella F. Sadat J.M. Torres-Moreno and K. Smaïli

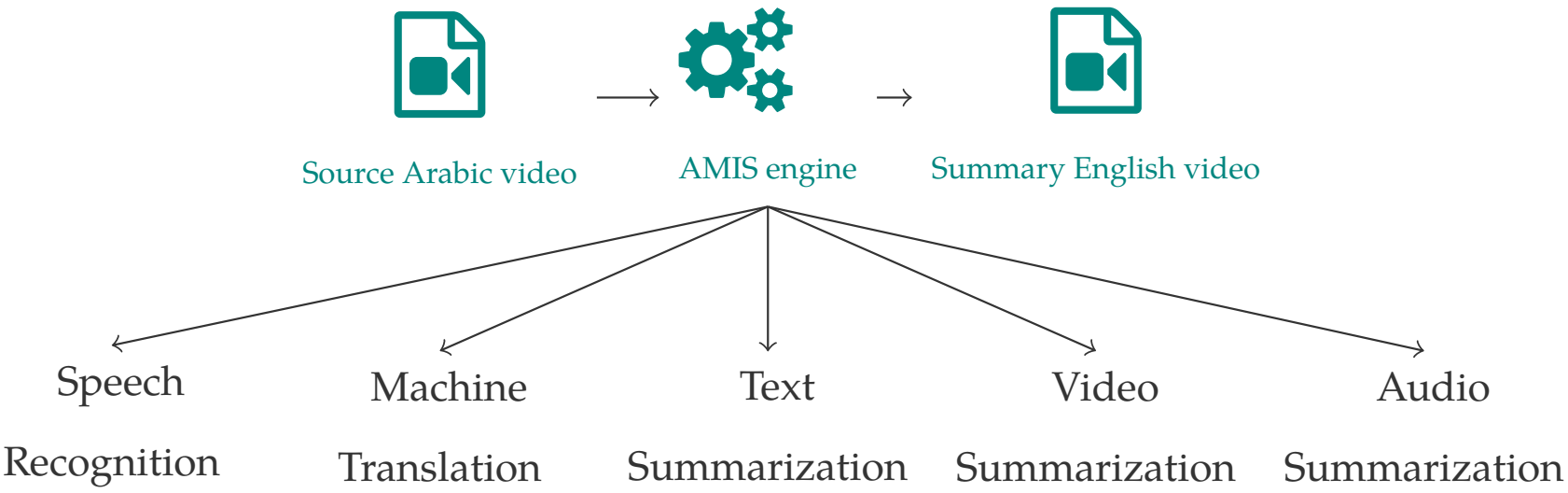
---

Nancy, France

16, 17 October 2019

Introduction: Context and objectives

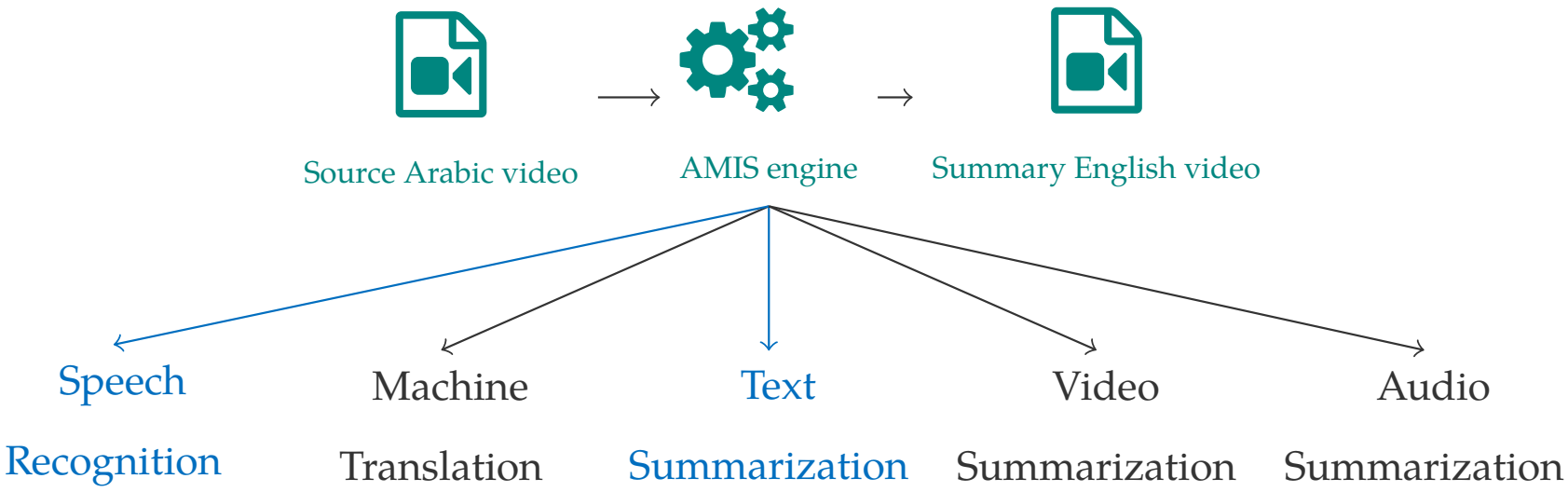
- How a user can access to the information which is expressed in a foreign language?
- Understanding a video in a foreign language is first step to answer this question.



- Develop and evaluate a system for automatic summarization of Arabic videos.

Introduction: Context and objectives

- How a user can access to the information which is expressed in a foreign language?
- Understanding a video in a foreign language is first step to answer this question.

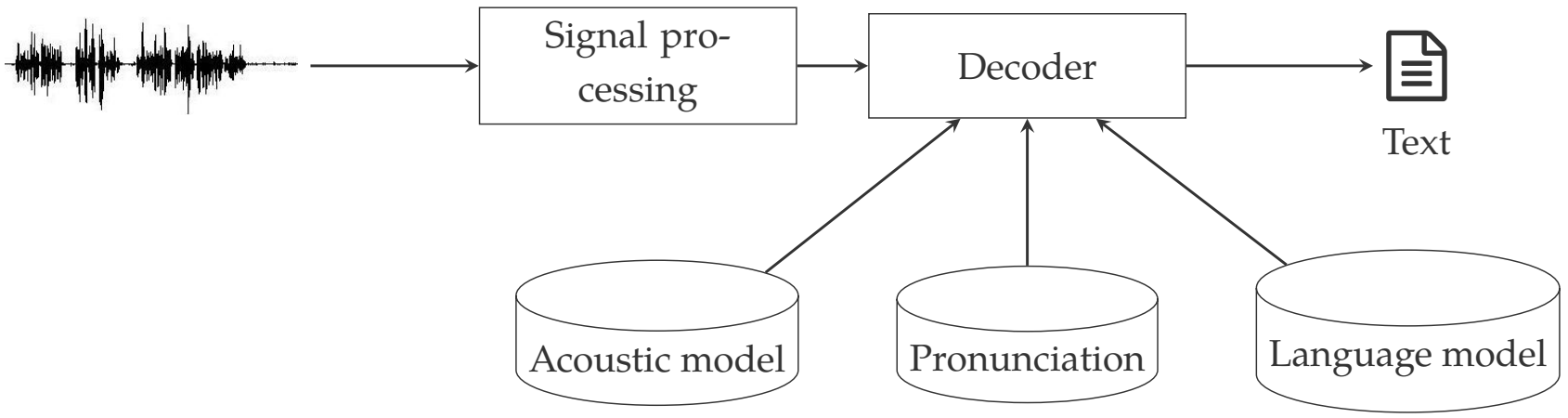


- Develop and evaluate a system for automatic summarization of Arabic videos.

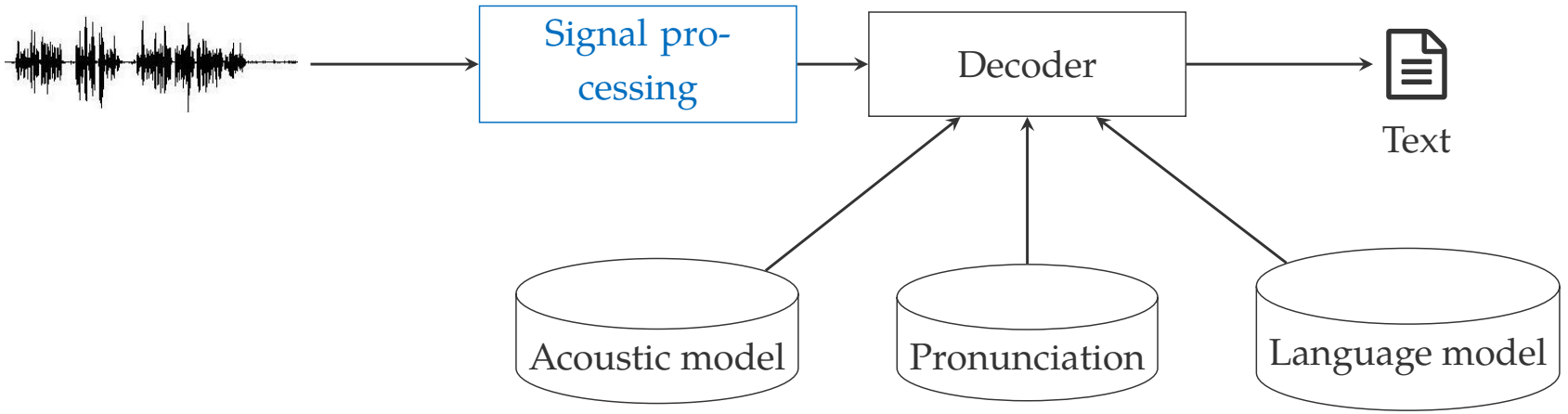
Automatic Speech Recognition -ASR-

Modern Standard Arabic and dialect cases

ASR: From the signal to the text

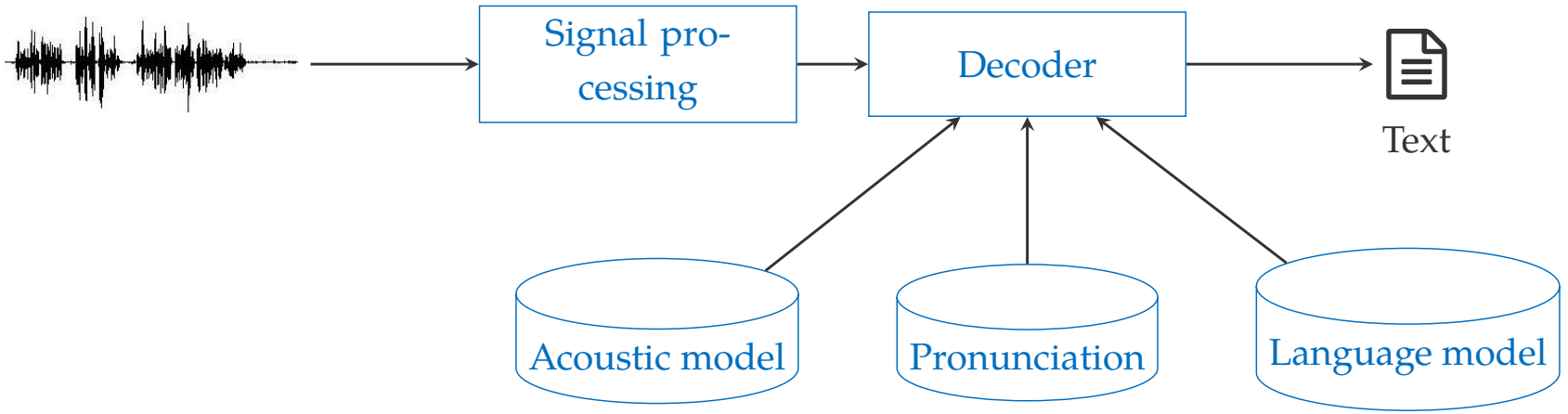


ASR: From the signal to the text



- Extract the acoustic features (MFCC, PLP ...).

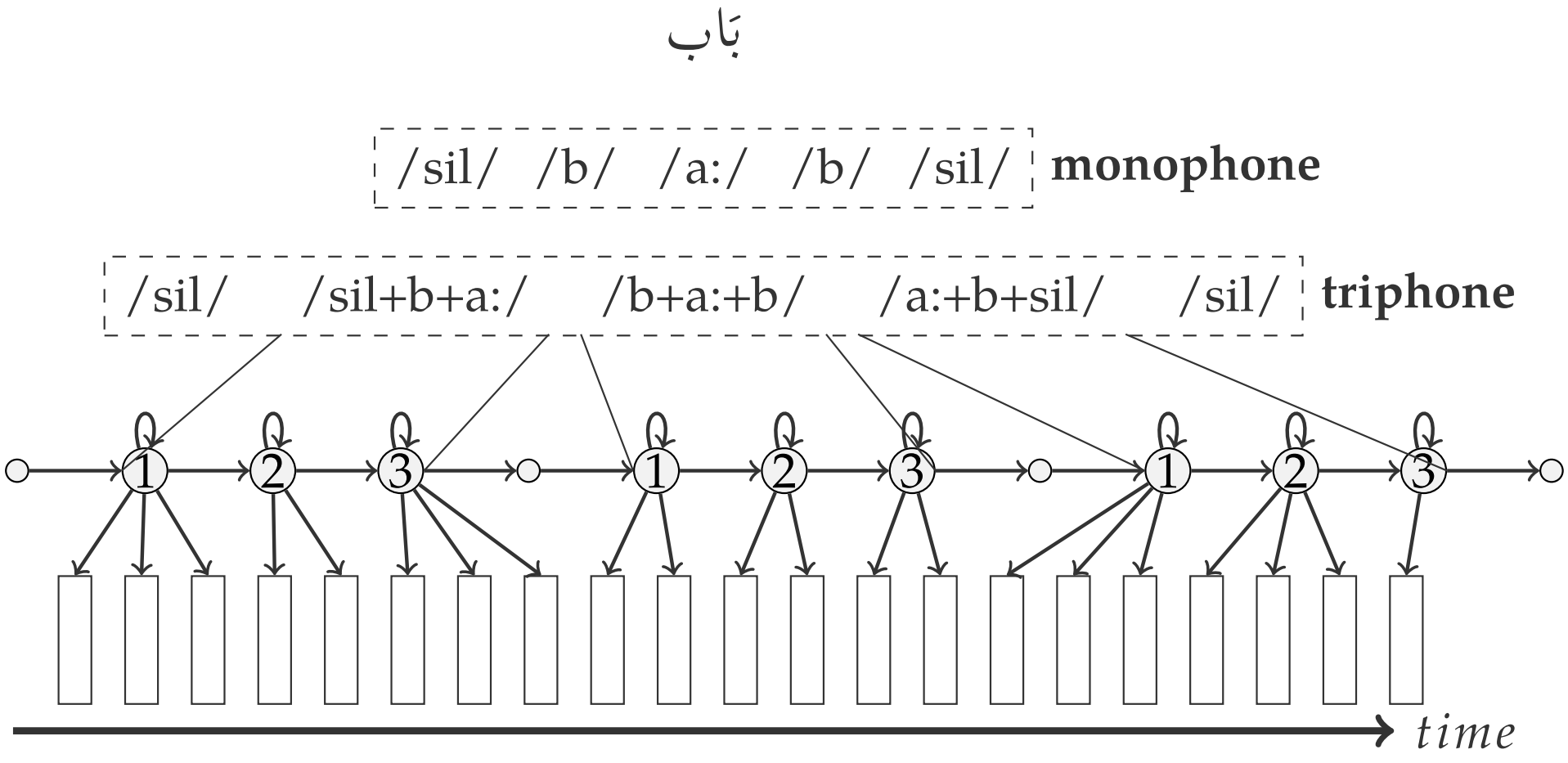
ASR: From the signal to the text



- Extract the acoustic features (MFCC, PLP ...).

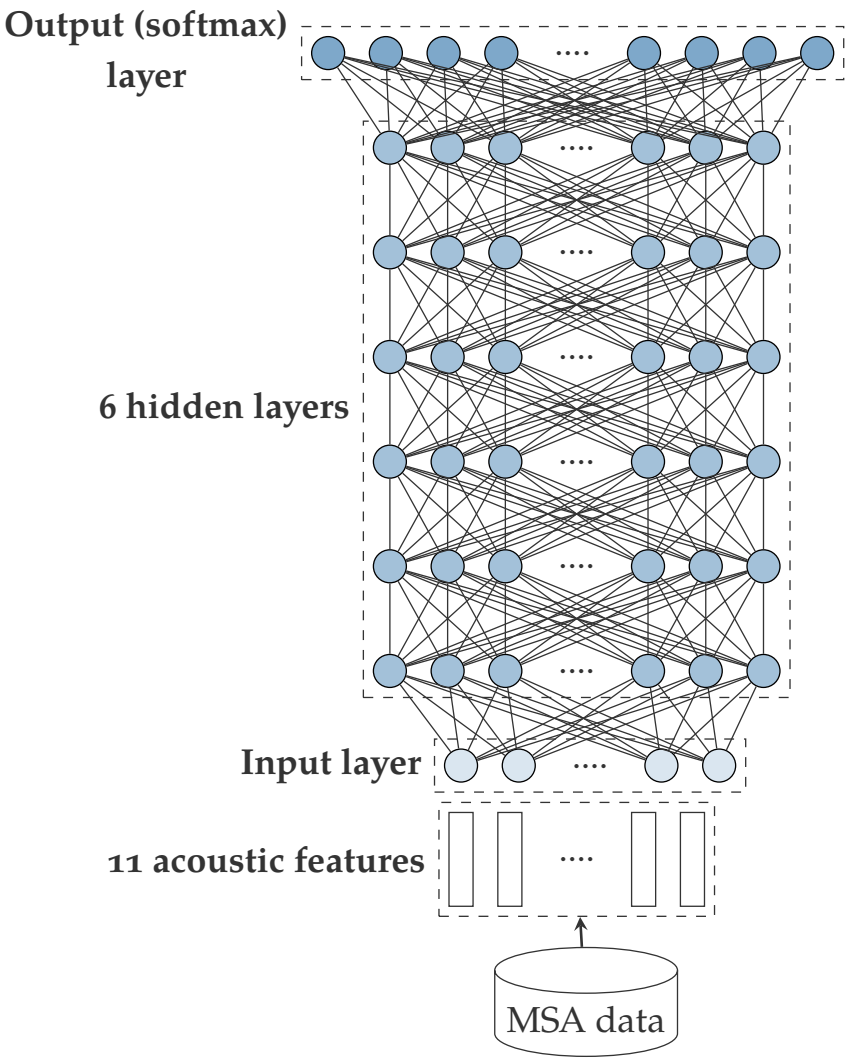
ASR: Modern Standard Arabic case

- Acoustic Modeling:
  - DNN-HMM model is used for the acoustic modeling.



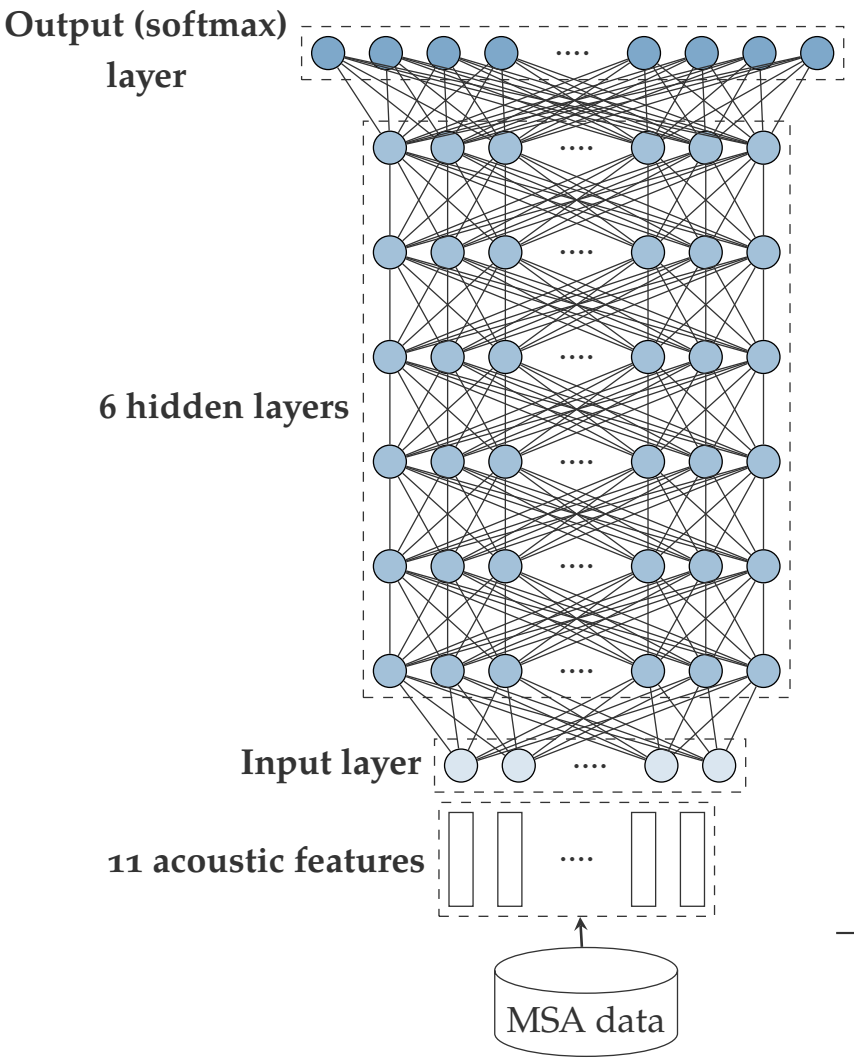
ASR: Modern Standard Arabic case

- Acoustic modeling: DNN-HMM acoustic modeling



ASR: Modern Standard Arabic case

- Acoustic modeling: DNN-HMM acoustic modeling

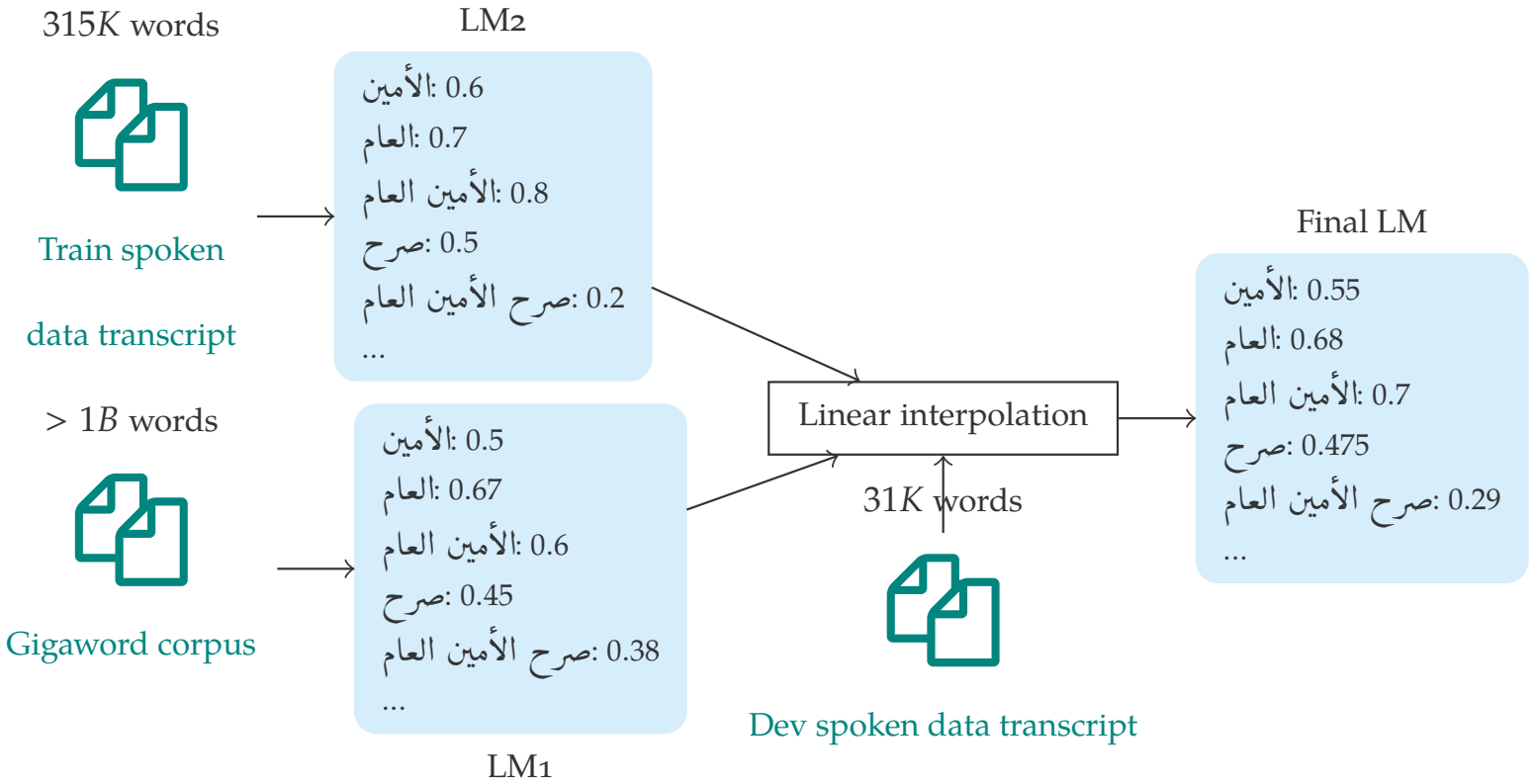


- 44 hours of MSA spoken data are used for training the neural network: Nemlar[a] and NetDC[b];
- 440-dimensional input layer (11*40-dimensional acoustic features);
- 6 hidden layers with 2048 nodes by layer;
- and 4264-dimensional output layer (number of HMM states).

[a] http://catalog.elra.info/product_info.php?p

[b] http://catalog.elra.info/product_info.php?products_i

- Language modeling:
  - n-gram model is used for the language modeling.



$$P(W) = \prod_{i=1}^{M} P(w_i | w_{i-1} \dots w_{i-n-1}) \quad (1)$$

$$P^{FinalLM}(W) = \lambda_1 P^{LM1}(W) + \lambda_2 P^{LM2}(W) \quad (2)$$

ASR: Modern Standard Arabic case

- Pronunciation modeling:
  - Select 100k most frequent words from the textual data.
  - Use an external lexicon[1] to generate pronunciation.

	#Words	**#Entries**
MSA	95K	485K

Table: Statistics about the MSA lexicon.

[1] http://alt.qcri.org/resources/msa-dictionary/

ASR: Algerian dialect case

- This Algerian dialect is highly impacted by the MSA and French language.
- The Algerian dialect is mainly spoken, there are no data to train the different model.

ASR: Algerian dialect case

- This Algerian dialect is highly impacted by the MSA and French language.
- The Algerian dialect is mainly spoken, there are no data to train the different model.
- Explore data that impact the Algerian dialect, namely MSA and French to enhance models for the dialect.

ASR: Algerian dialect case

- Textual data collection:
  - two corpora containing Algerian dialects are constituted: PADIC[2] and CALYOU[3] corpora.

Corpus	#Words	#Unique words
CALYOU	10M	512k
PADIC	25k	6.6K

Table: Statistics about trxtual data.

[2]K. Meftouh, S Harrat, and Kamel Smaïli. "PADIC: extension and new experiments". In: *7th International Conference on Advanced Technologies ICAT*. Antalya, Turkey, Apr. 2018. URL: https://hal.archives-ouvertes.fr/hal-01718858.

[3]karima Abidi, Mohamed amine Menacer, and Kamel Smaili. "CALYOU: A Comparable Spoken Algerian Corpus Harvested from YouTube". In: *18th Annual Conference of the International Communication Association (Interspeech)*. 2017.

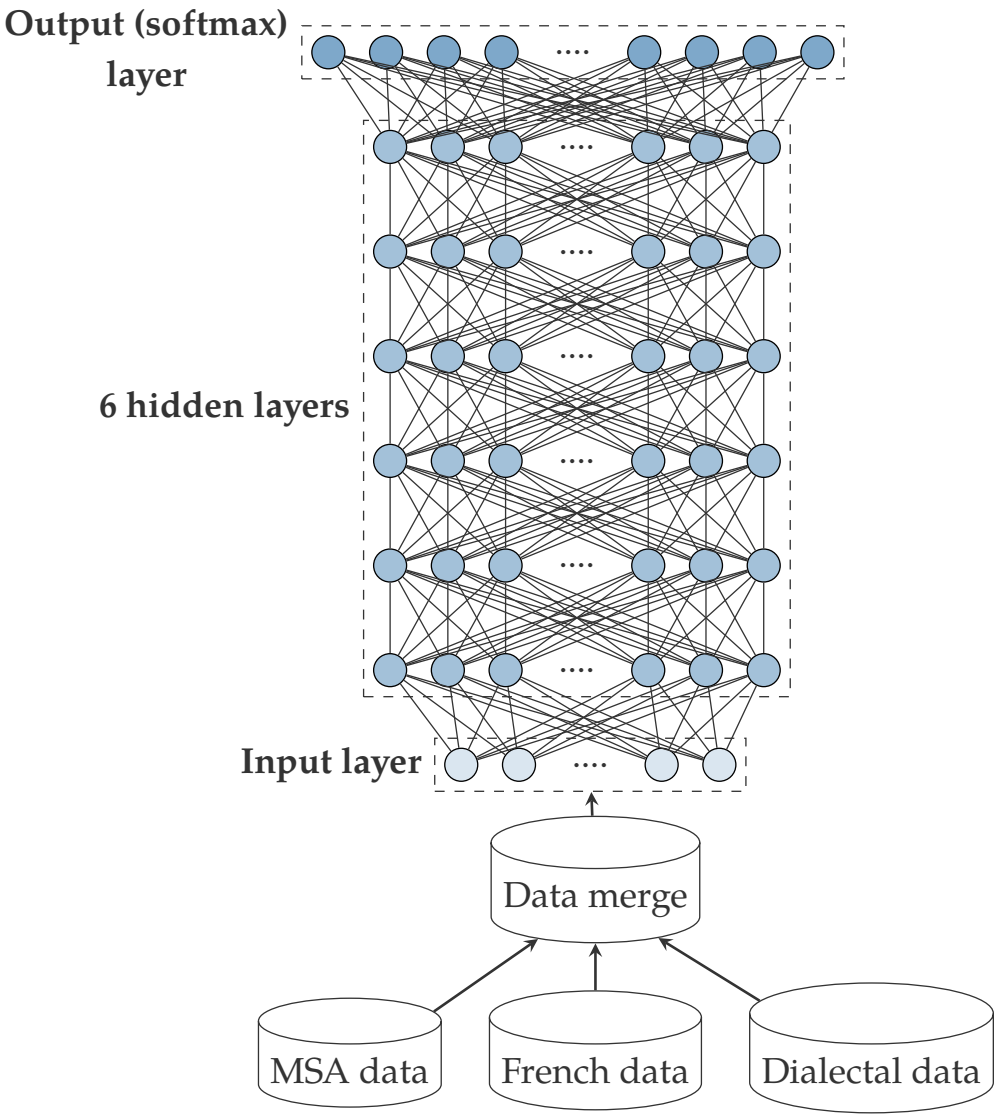
- Spoken data:
  - The aligned dialectal spoken corpus is created by having native Algerian people reading 4.6k sentences extracted from PADIC and CALYOU corpora.

Subset	Dur	Female spkrs	Male spkrs	Total spkrs
Train	240 min	1	3	4
Dev	40 min	1	1	2
Test	75 min	1	2	3

Table: Some characteristics of the dialectal corpus.

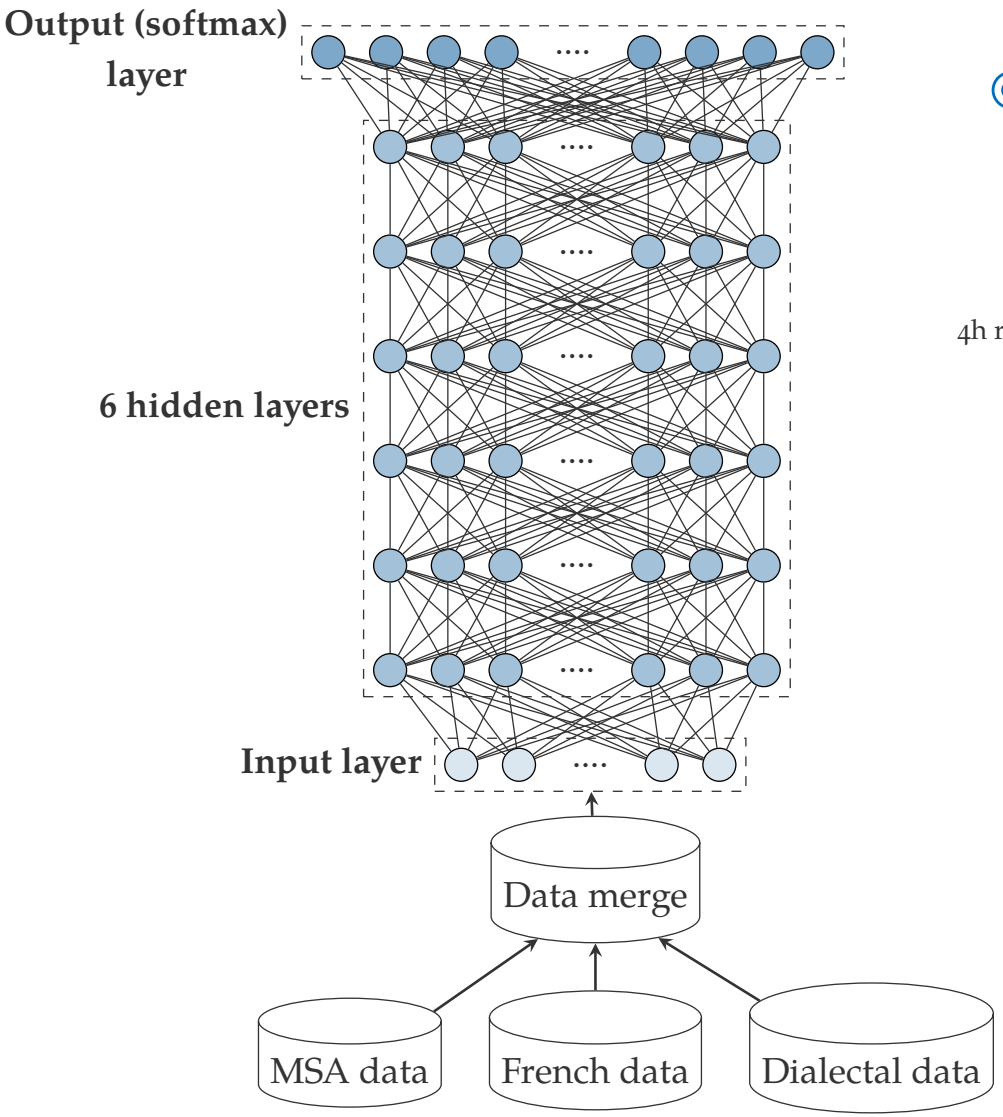
ASR: Algerian dialect case

- Acoustic modeling:
  - The dialectal corpus is quite small to train a robust AM.

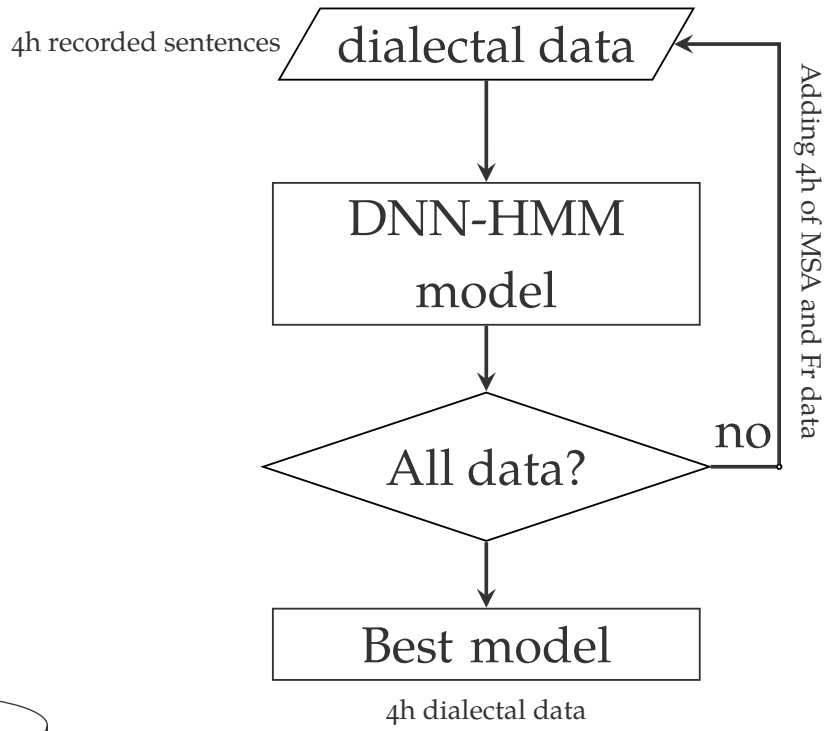


ASR: Algerian dialect case

- Acoustic modeling:
  - The dialectal corpus is quite small to train a robust AM.

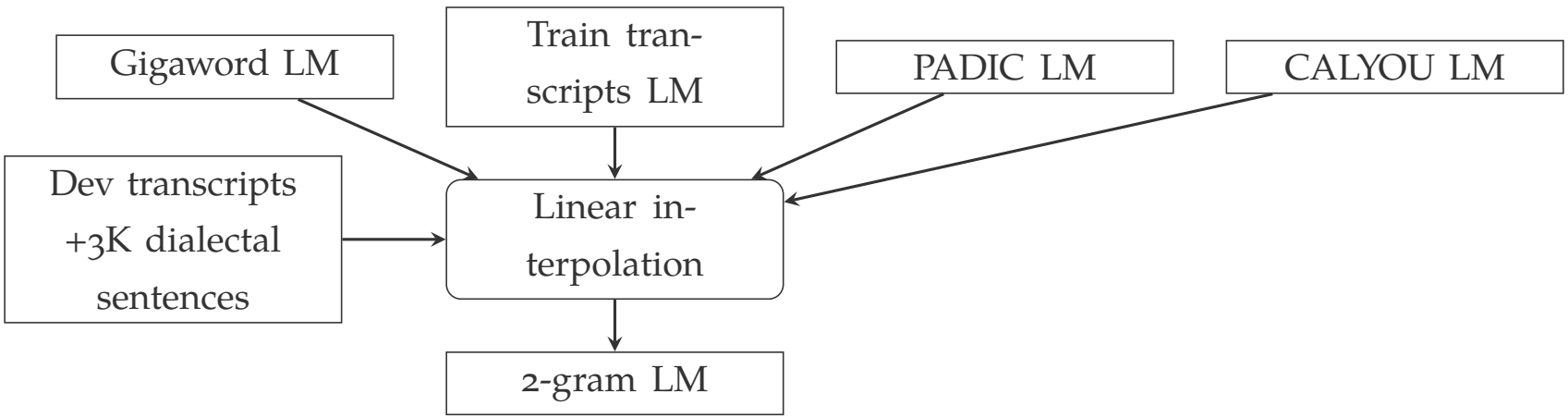


- The amount of data is optimized iteratively on the dialectal Dev corpus.



ASR: Algerian dialect case

- Language modeling:
  - The LM is a linear interpolation of 4 LMs.



ASR: Algerian dialect case

- Pronunciation modeling:
  - Adapt the approach proposed in[4] to generate the pronunciation of dialectal words.

Corpus	#Words	#Entries
MSA	95K	485K
CALYOU	50K	50K
PADIC	6.6K	6.6K
Total	123K	538K

Table: Statistics about lexicons.

[4]Salima Harrat et al. "Grapheme to phoneme conversion-an arabic dialect case". In: *Spoken Language Technologies for Under-resourced Languages*. 2014.

ASR: Results

- The test is carried out on the 75 min of the dialectal data and 5 hours of MSA data:

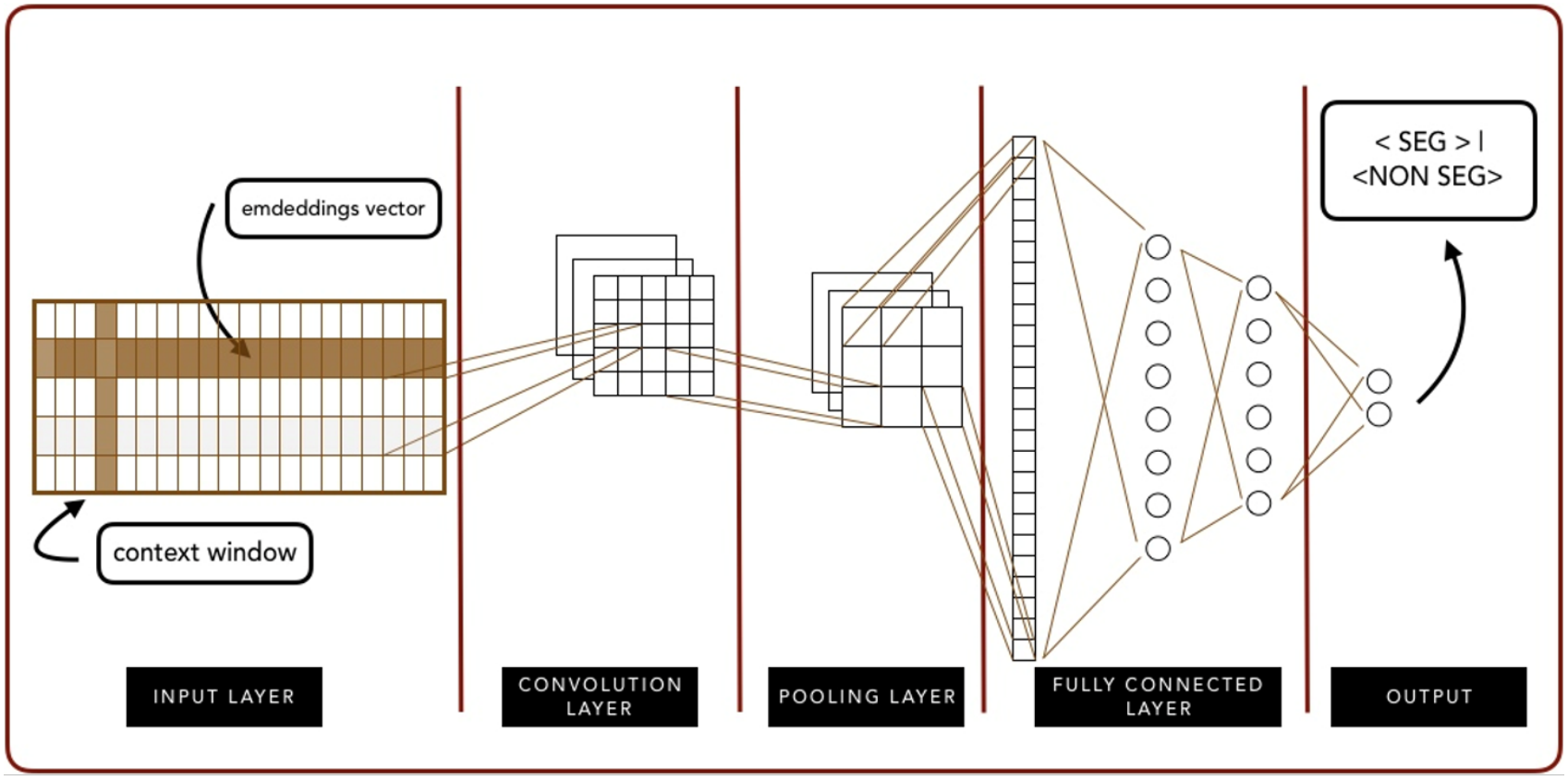
System	AM	LM	Lex	WER_dial (%)	WER_MSA (%)
ASR-MSA	MSA	MSA	MSA	78.5	14.02
S_1	4h dial	MSA+dial	MSA+dial	40	/
S_2	MSA+Fr+dial	MSA+dial	MSA+dial	37.7	/

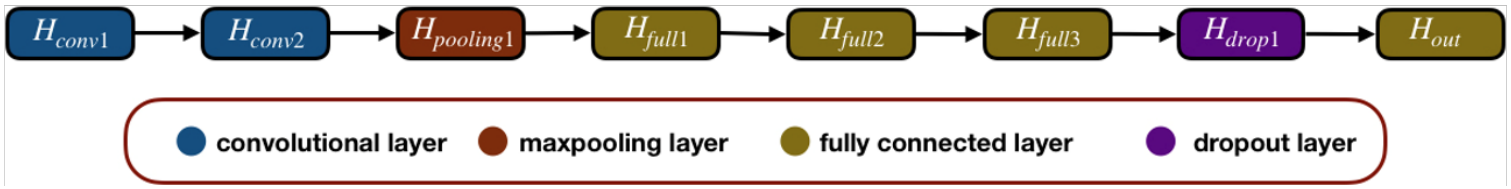
Table: Performance of the ASR systems on the Test dialectal corpus.

Automatic text summarization

Sentence Boundary Detection

Sentence Boundary Detection: Architecture





Sentence Boundary Detection: Results

- The CNN is trained on 70M words subset extracted from the Gigaword corpus.
- The evaluation is carried out on 10.5M samples.

class	Precision	Recall	F1-score
<SEG>	0.797	0.612	0.684
<NO SEG>	0.972	0.989	0.98

Table: Sentence Boundary Detection performance.

Automatic text summarization

Automatic text summarization

Automatic text summarization: ARTEX algorithm[5]

Document preprocessing The text is represented in a suitable space model.

Global topic vector An average document vector is built.

Lexical weight A lexical vector is built for each sentence.

Sentence scoring A score for each sentence is calculated using their proximity with the global topic vector and their lexical weight.

$$score(s_i) = (\overrightarrow{s} \times \overrightarrow{b}) \times \overrightarrow{a} = \frac{1}{NP}(\sum_j s_{i,j} \times b_j) \times a_i \qquad (3)$$

Sentence selection The summary is generated concatenating the sentences with the highest scores following their order in the original document.

[5]Juan-Manuel Torres-Moreno. "Artex is AnotheR TEXt summarizer". In: *CoRR* abs/1210.3312 (2012). arXiv: 1210.3312. URL:

Tests ans results

Evaluation

Evaluation: video corpus

- French, English and Arabic videos are collected according to a set of controversial Twitter Hashtags such as #سوريا, #حقوق_المرأة.
- More than 1.5K Arabic videos (>100h) are collected. they come from channels such as AlArabiya, France24, EchoroukTV, EnnaharTV, BBC, etc.

Count	**Value**
Videos	27
Summary per Video	3
Channel TV	3
Evaluators	3
Size of the shortest summary (in words)	52
Size of the longest summary (in words)	394

Table: Some figures concerning the subjective evaluation.

Evaluation: Subjective evaluation, MSA case

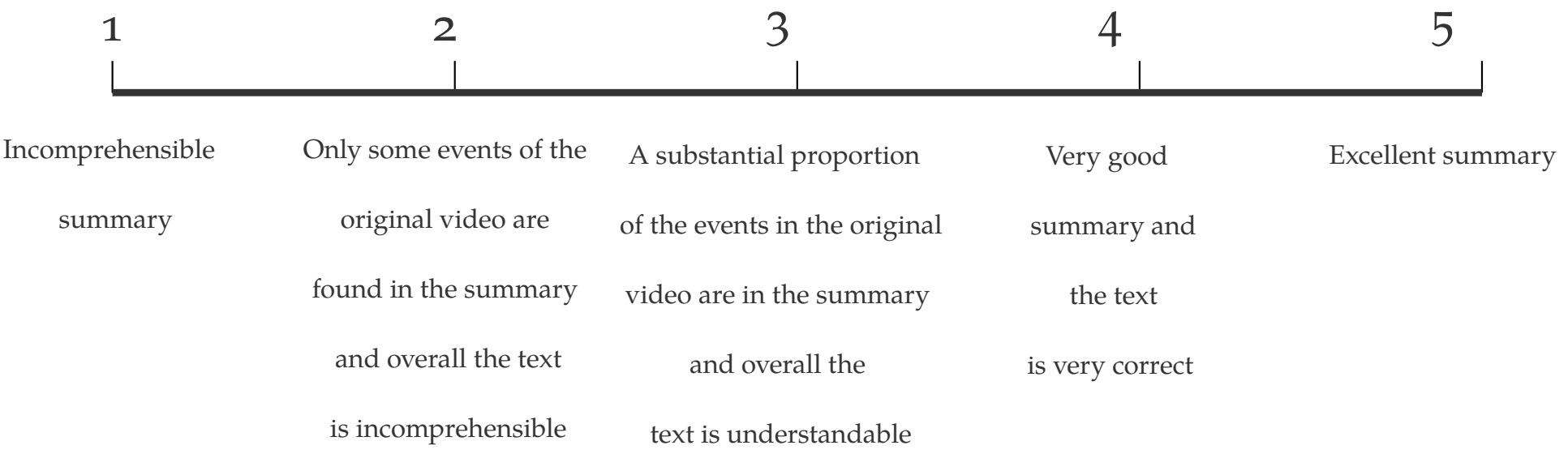


Figure: Rating scale for the automatic summarization system assessment.

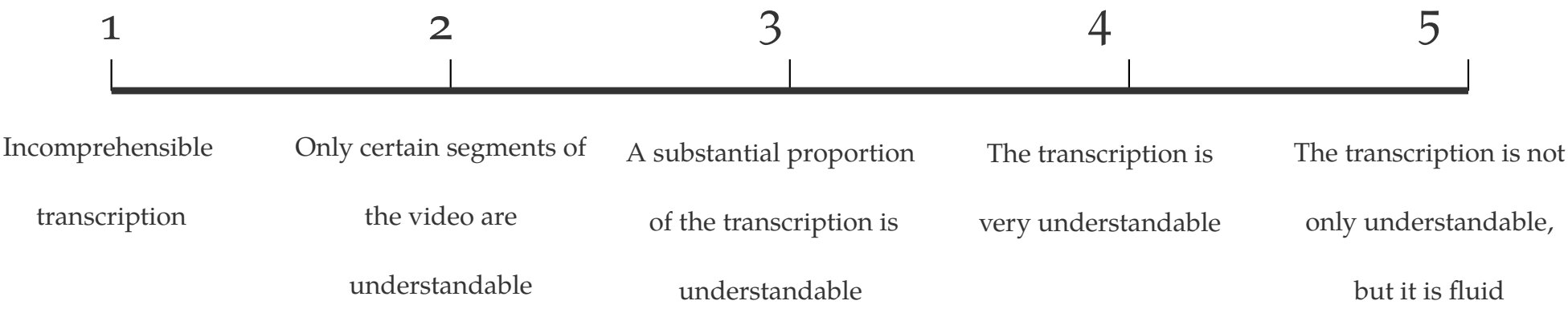


Figure: Rating scale for the automatic speech recognition system assessment.

Evaluation: Subjective evaluation, MSA case

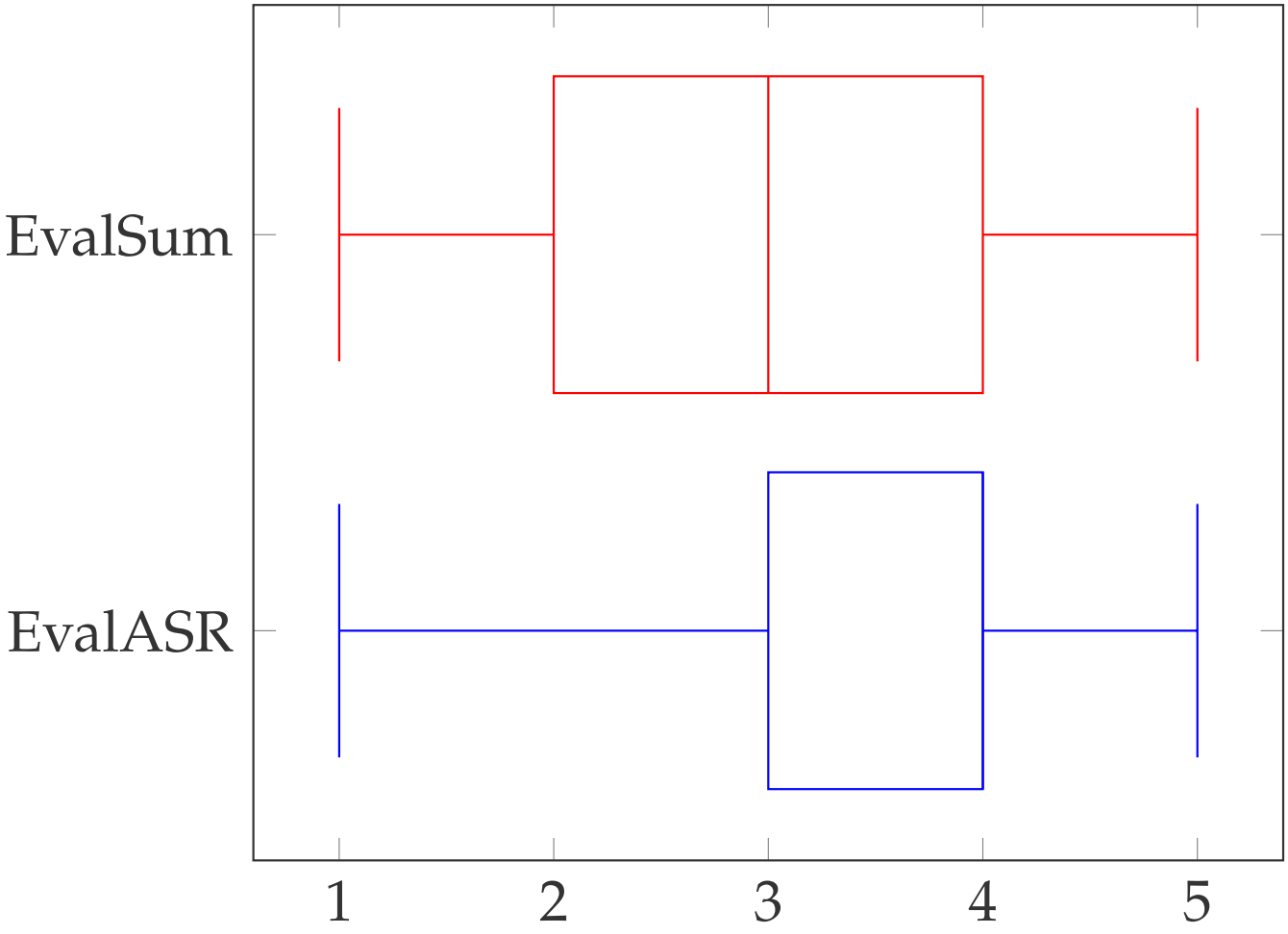


Figure: The Box plot corresponding to the subjective evaluation of the Arabic ASR and the automatic summarization systems on MSA data.

Evaluation: Subjective evaluation, Algerian dialect case

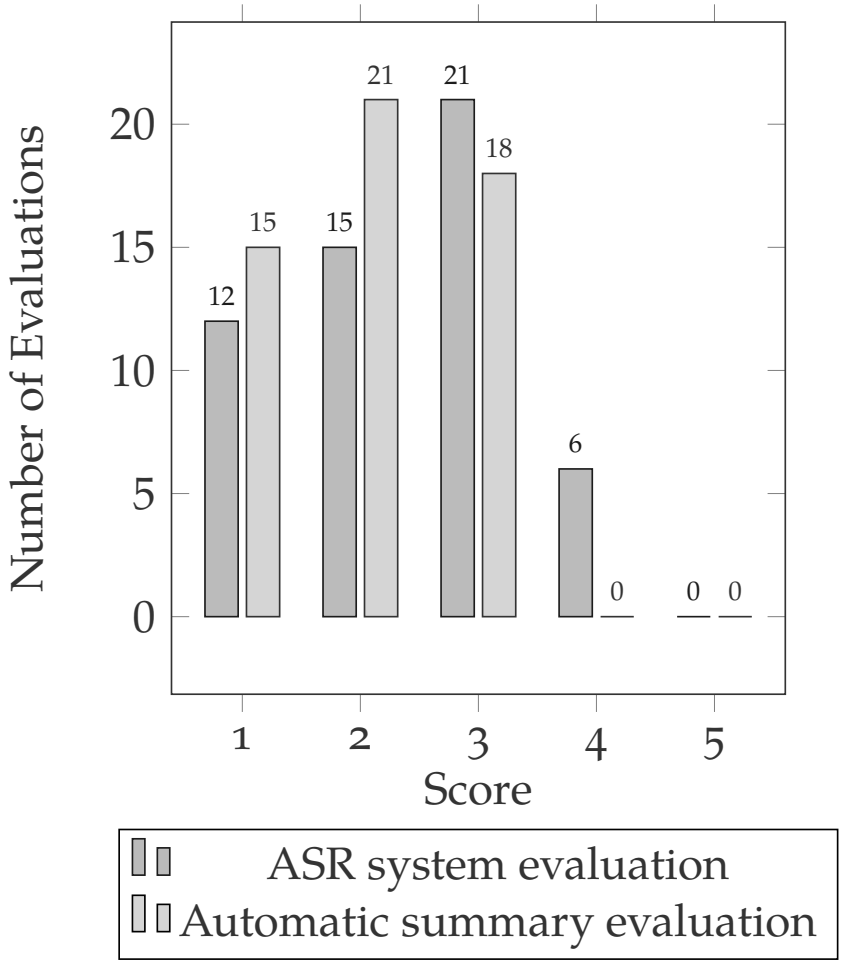


Figure: The number of responses for each score of the subjective assessment of dialectal data with MSA-ASR system.

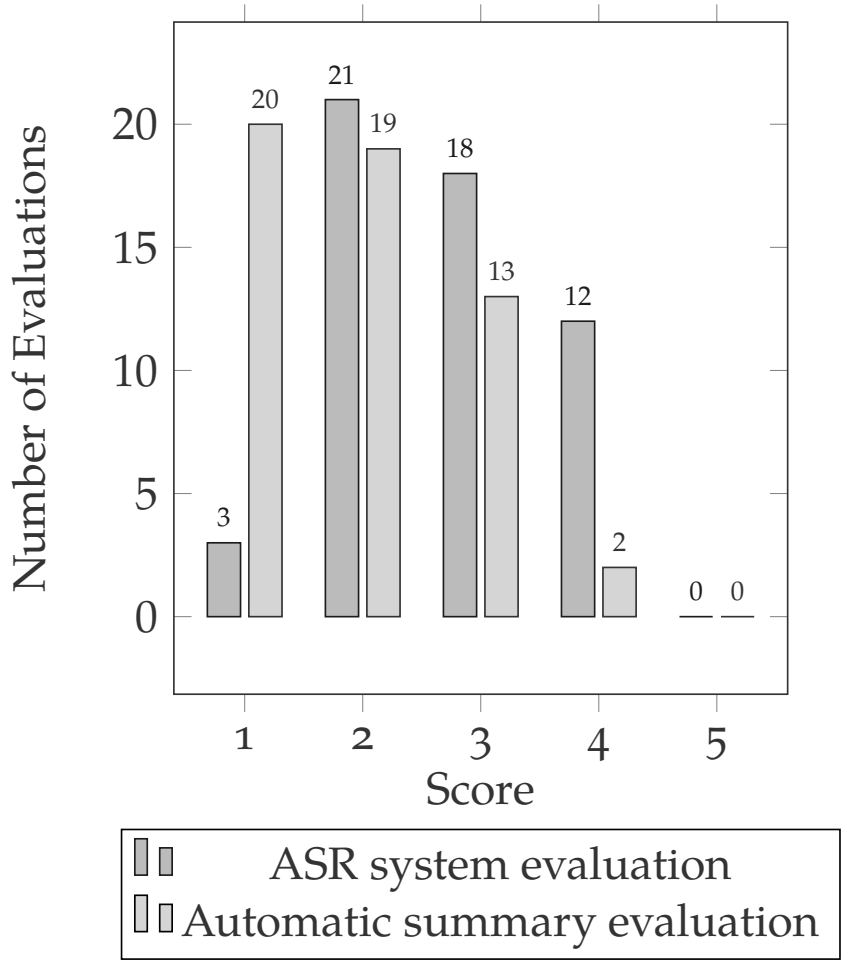


Figure: The number of responses for each score of the subjective assessment of dialectal data with the adapted ASR system.

Evaluation: Factors impacting summary

- What is the relationship between the scores of the summary $EvalSum$ and:
  - the number of words ($ASRWord$);
  - the score of the ASR system ($ASRScore$);
  - and the number of words of the summary ($SumWord$).
- Use the multiple linear regression through the coefficient of determination (R^2).
- On our data-set of 243 examples, $R^2 = 0.310$, this indicates that 31% of the dispersion is explained by the regression model.

Evaluation: Factors impacting summary

- $H_0 : a_1 = a_2 = a_3 = 0$ and $H1$ at least one of the a_i is different from 0.

$$F = \frac{\frac{R^2}{p}}{\frac{1-R^2}{n-p-1}} \qquad (4)$$

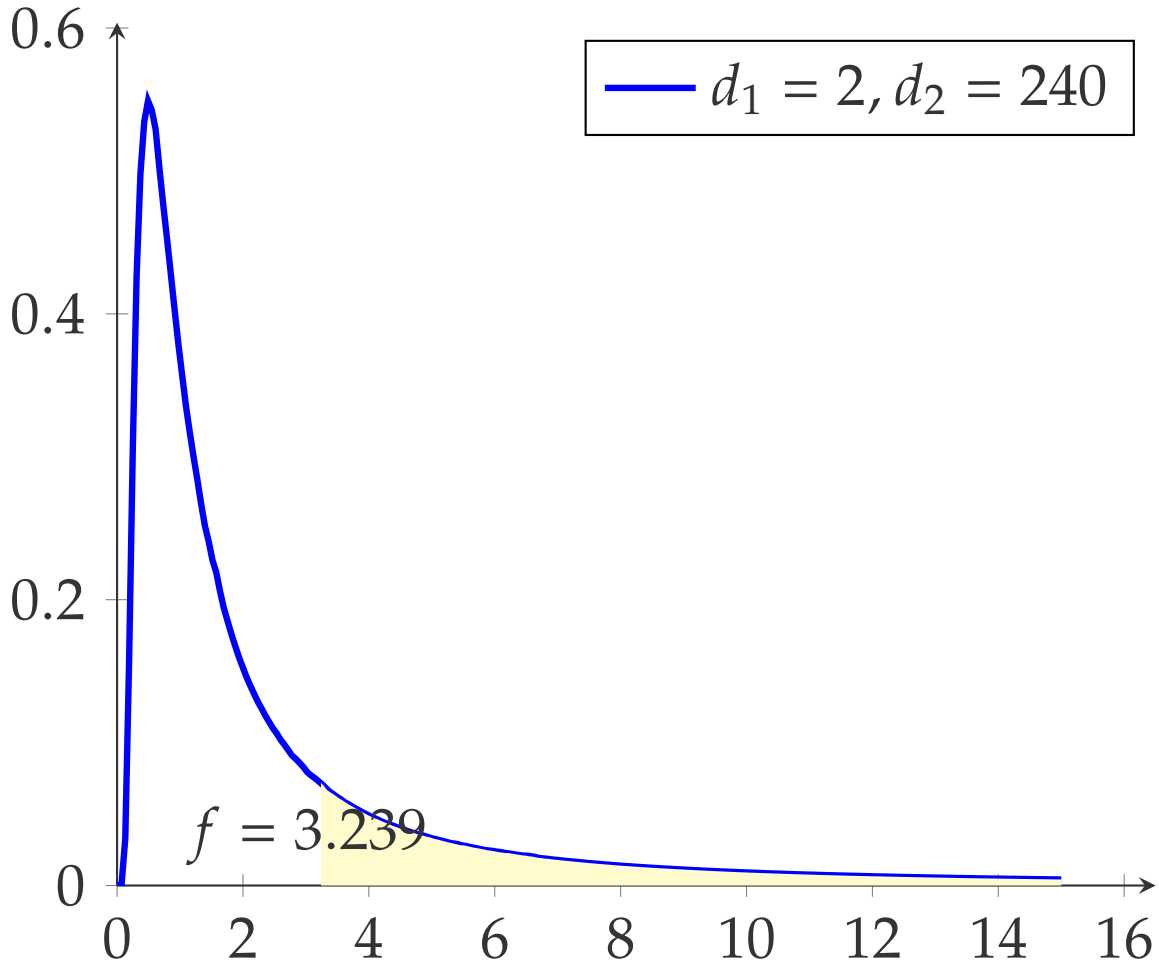
$R^2 = 0.31$

$p = 2$ #dependent variables-1

$n = 243$ #samples

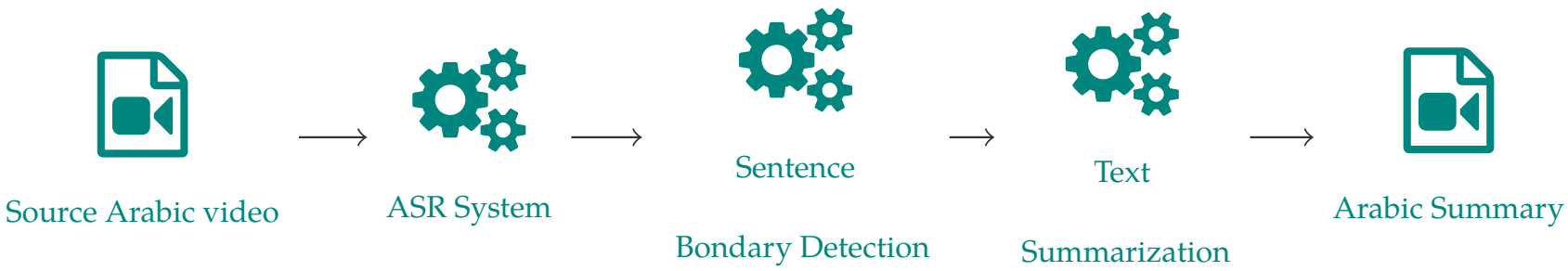
$F = 35.899$

$F > F_{0.95}(2, 240)$



Conclusion

- Describe the development and the evaluation of an automatic video summarization system.



- The ASR system was developed for MSA and adapted for the Algerian dialect.
- Each component performs well separately.
- Several parameters impact the summary, namely the number of words in the original/summarized video and the output of the ASR system.

Thank you

for your attention

Questions?