Evaluation of Improved Components of AMIS Project for Speech Recognition, Machine Translation and Video/Audio/Text Summarization

Aritz Badiola, Amaia Méndez Zorrilla, García-Zapirain Begoña, Michal Grega, Mikolaj Leszczuk, Kamel Smaïli

Evaluation of Improved Components of AMIS Project for Speech Recognition, Machine Translation and Video/Audio/Text Summarization

Aritz Badiola[1], Amaia Méndez Zorrilla[1], Begoña García-Zapirain Soto[1], Michał Grega[2], Mikołaj Leszczuk[2][0000-0001-9123-1039], and Kamel Smaïli[3]

[1] University of Deusto, Bilbao, Spain
[2] AGH University of Science and Technology, Kraków, Poland
[3] University of Lorraine, Nancy, France
leszczuk@agh.edu.pl

Abstract. To evaluate a system that automatically summarizes video files (image and audio) and text, how the system works, and the quality of the results should be considered. With this objective, the authors have performed two types of evaluation: objective and subjective. The actual assessment is performed mainly automatically, while the individual assessment is based directly on the opinion of people, who evaluate the system by answering a set of questions, which are then processed to obtain the targeted conclusions. One of the purposes of the described research is to try to narrow the space of possible summarization scenarios. Meanwhile, in the light of individual results obtained, the researchers cannot unambiguously indicate one single scenario, recommended as the only one for further development. However, the researchers can state with certainty that the new development of scene 1, which has received many negative evaluations among professionals, should be discontinued. Considering the results of the set of questions about the quality of the complete system, the end-users have evaluated the scenario 3, and they think that the quality is excellent, obtaining results over 70% on a scale of 0 to 100.

Keywords: Subjective Evaluation, Summarization system, Summarized videos, AMIS project.

1 Introduction

Nowadays, information is widely available in different media: TV, social networks, newspapers, etc. in foreign languages. When the video does not necessitate any understanding, there does not have great difficulty. In the opposite, when the information requires understanding the word, a human being is limited in terms of mastering a foreign language. One of the main objectives of the AMIS project is to make available a system, helping people to understand the content of a source video by presenting its main ideas in a target understandable language. Some scholars believe that the best way to do that is to summaries the video for having access to the essential information.

AMIS focuses on the most relevant information by synthesizing it and by translating it to the user if necessary.
Several capabilities are necessary to achieve this objective: video summarization, automatic speech recognition, machine translation, text summarization, etc.
In this article, we will present the first results of the subjective evaluation of the proposed architectures of AMIS [1], and personal assessment of the whole system with videos in Arabic.
The research team decided that it is a waste of resources to develop several scenarios at the same time, so the researchers should determine what scene will be used for further development.
The research team decided that the language of the recorded source video sentences would be Arabic and French.

2 Objective

As presented in the previous section, the problem to be solved is the development of an evaluation procedure for the developed video summarizer in the project AMIS. The development of this evaluation procedure is considered necessary, as there is not a standardized way of evaluating the summarizer systems, so, developing a tested and complete evaluation procedure for video summarizes is the objective of this project.

To develop this evaluation procedure, it has been used the already developed video summarization system in the project AMIS as a use case. The used method to develop this evaluation procedure is presented in the next sections, including the subjective evaluation procedure, the actual evaluation procedure and the obtained results with each system.

3 Subjective Evaluation

In this section, the material and methods used for the different evaluations made in the AMIS project are described.

3.1 Methodology

Methods for Subjective Video Evaluation of the Generated Scenarios. To evaluate the three scenarios generated by AMIS project methods, a set of 9 video sequences manually taken to represent diversified content have been used. From the source video sequences, 27 sequences were subsequently generated (for each source sequence, three series summarized, in scenarios 1, 3 and 4). In the next step, file names were randomized, in such a way that the outsider could only learn the source sequence number, but not the amount of the summary scenario used for the sequence summarization. The research team distributed the thus prepared video sequence packet within the project consortium and students. Members of the AMIS consortium was asked to rank, for each of the nine source video sequences, generated three summaries (from the best to the worst) – of course, without knowing the assignment of a specific video summary file

to the number of the summary scenario. We addressed two groups of respondents: students (pre-teenagers) and professionals (scientists, academic teachers, technical workers – generally adults, working full-time). Sixty-two people (51 students and 11 professionals) responded to the call, generating a total of 558 sets of data, fed to the following analyses.

First, collective analysis of all answers was carried out. Then (for reasons that we will give later), separately in the groups of students and professionals.

Methods for Subjective Video Evaluation of the Generated Scenarios. After selecting the best scenario for video summarization considering the obtained results from the evaluation of the different scenarios, a subjective evaluation of the summaries generated but that scenario has been performed, to evaluate the final system performance.

The only applied inclusion criteria when selecting the users for the evaluation is that the user must be a native speaker of the language spoken in the video and have a right level of English, at least a B2 level to understand the questions of the questionnaire.

The age of most of the users that took part in the evaluation was between 18 and 35 years old, and their mother language was French or Arabic, and their nationality French or Arab. The maximum level of education of most of them was a bachelor's degree, but there were some with an elementary school, high school, master or doctoral degree.

As said, the objective of this evaluation process is to analyze how good the summarization system performs with Arabic and French videos. With this purpose, an evaluation section has been created on the web page of the project, where the users can register themselves and evaluate different videos.

The evaluation section of the web page has been structured in the following way:

- An introduction, explaining to the user the purpose of the evaluation and the steps to do it.
- The video summarization generated by the system.
- Questionnaire part:
    - Five questions about the critical points of the original video to ensure that the summary video includes the main ideas and the user can get these ideas from the summary. Three possible answers per question are provided to the user.
    - Two generic questions about the quality of the video summary. The possible answers are from 0 to 4, "Not fulfilled", "Fair", "Good", "Very good", "Excellent". The questions are the following ones:
        - "Is the summary understandable?".
        - "Doesn't the video contain any part out of context, or it does not affect the main expressed ideas?".

3.2 Results of Subjective Evaluation of the Different Scenarios

Subjective Evaluation of All People. The first approach analyzed the percentage distribution of "winners" (see Fig. 1). By winners, we mean scenarios indicated as the best

during the evaluation. The consortium members expected a definite answer at this stage, but this did not happen.

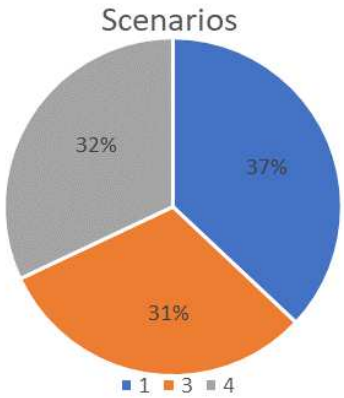

Fig. 1. Subjective evaluation winners.

The result significantly surprised the experimenters. As one can see, the differences are negligible. All scenarios achieve a similar effect, dividing the "pie" by approximately 1⁄3. Differences between individual scenes are in practice different percentage points, which does not allow to draw firm conclusions.

Since the research team could not unequivocally set the best scenario, the researchers asked a research question differently: is it possible to at least indicate the worst-case scenario (or the worst scenarios)? So, reject this scenario (these scenarios)? Therefore, such an analysis of "losers" was carried out (see Fig. 2).

Nevertheless, again, as one can see, the differences are rather negligible. All scenarios achieve a similar result, dividing the "pie" by approximately 1⁄3. It was somewhat surprising because, in the subjective opinion of the project members, the differences between the scenarios were quite visible and significant.

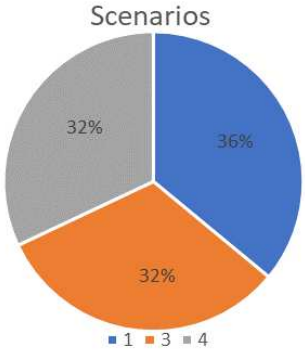

Fig. 2. Subjective evaluation of losers.

Subjective Evaluation of Students. In the case of the group of students, the result was still not decisive. Practically repeated the scheme of results previously known during the analysis of the entire group of respondents.

The analysis of winners (see Fig. 3) have not provided precise results. All scenarios achieve a similar effect, dividing the "pie" by approximately 1⁄3. So, students pointed to each scene as the best, more or less in the same number of cases.

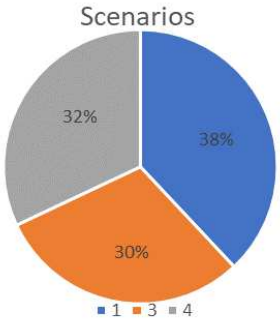

Fig. 3. Subjective evaluation winners among students.

Moreover, similarly, the analysis of losers (see Fig. 4) was not decisive. All scenarios achieve a similar result, dividing the "pie" by approximately 1⁄3. So, students pointed to each scene as the worst, more or less in the same number of cases.

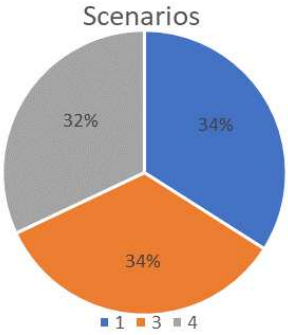

Fig. 4. Subjective evaluation losers among students.

Subjective Evaluation of Professionals. In the case of a group of professionals, the result still turned out to be very interesting. As for the winners, he was still not decisive (see Fig. 5). All scenarios achieve a similar result, dividing the "pie" by approximately 1⁄3. So, professionals pointed to each scene as the best, more or less in the same number of cases.

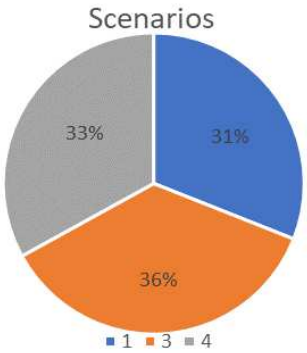

Fig. 5. Subjective evaluation winners among professionals.

However, regarding the losers, this time, it is evident that the testers indicated scenario number 1 as the worst in almost 1⁄2 of cases. The other two scenarios (3 and 4) divide the "pie" more or less equally, with no visible indication of any of the remaining scenes (see Figure 6). However, at least, this time, differences between individual scenario one and the two remaining ones are in practice over a dozen percentage points, which finally does allow to draw firm conclusions.

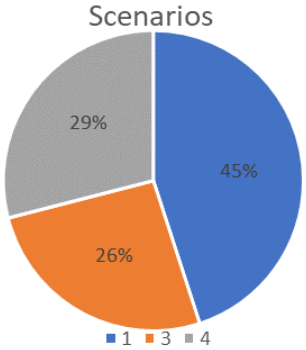

Fig. 6. Subjective evaluation losers among professionals.

Why are these two groups so different in terms of pointing out the worst-case scenario? It is not known precisely. Indeed, professionals were well motivated to perform the test reliably: either they worked on the project themselves, or the persons working on the project personally asked other professionals to do the test. Meanwhile, students did not have a similar motivation. At the same time, it is worth noting that differences can also result from one or a combination of a more significant number of entirely different, partly independent reasons, such as age, education, knowledge or experience.

Conclusions. As mentioned, the purpose of the described research was to try to narrow the space of possible scenarios. Therefore, in the light of individual results obtained, the researchers cannot unambiguously indicate one single scene, recommended as the only one for further development. However, the researchers can state with certainty that the new development of scenario 1, which has received many negative evaluations among professionals, should be discontinued.

Along with all the possible scenarios, it has been selected the scene 3 for the final system evaluation. It is true that in most of the assessment of the situations performed by all the participants, the results of the three scenarios are similar. However, in the case of the evaluations by the professionals, the results of scene 3 are prominently better than the ones obtained for the rest of the scenarios.

More information about the methods using which these results were obtained has been provided in [1].

3.3 Results of the Subjective Comprehension-Evaluation of Scenario Three Videos

This section includes the obtained results from the tests performed by the users about the summaries produced by the selected best scenario from the previously performed evaluation, which is ranged from 0 to 100. This score is the result of the answers to the five specific questions about the main ideas of the video. Each correct answer is 20 points, so, if the five are correct, then the user will get 100 points (being acceptable from 70). It is included another information that is considered attractive too, such as the number of videos, users, answers, standard deviation and the results of the generic questions.

There is not too much work performed about the evaluation of this type of systems. Still, there is a perfect example to understand how the review should be approached inside the publication Multimedia summarization using social media content [4]. In this paper, the evaluators see the multimedia content. Then they are asked to summaries all the content in two distinct triples of keywords and two different sets of summaries, each one containing 15, 25 and 50 images from the list of candidates computed by the algorithm.

In the case of the system being analyzed in this paper, it has been considered to apply a more complex and flexible approach for the subjective evaluation by users, as established by [5] as extrinsic evaluation procedure, where the quality of the summarization based on how it affects the completion of some other tasks, such as question-answering and comprehension tasks. Considering that the summary task more complicated due to the multilingual content and the duration of the videos and that the final objective of summarization is giving to the user the necessary information to understand the original content, it has been considered the extrinsic Q&A method as the best one for this case, with some extra details that will be explained in the next paragraphs. Other papers such as the challenging task of summary evaluation: an overview [6] analyses different ways of evaluating summarization systems too, including questionnaires answered by users about the quality and the content of the summaries about the original content, in a similar way it is done into this case.

Some users from different cultures and nationalities summaries the original videos in a set of 5 questions that include the main ideas of the video, so, then, other users can answer the questions only watching the summarized videos, evaluating in this way the quality of the system.

Arabic Video Summarization System Evaluation Result.

Number of videos	Number of users	Number of answers	AVG score	AVG SD	GQ_AVG[1]
8	20	35	80.56	15.71	2.46 (between good and very good)

Fig. 7. Overall results of the evaluation of the Arabic video summarization system.

[1] Generic Question Average: as explained before, there are 2 generic questions about the quality of the video summary included in the questionnaire, which range from 0 to 4: "Not fulfilled", "Fair", "Good", "Very good", "Excellent".

Video ID	Number of users	AVG	SD	GQ_AVG
1	4	60	28.28	2.5
2	2	66.66	11.55	2.16
40	5	80	12.65	1.58
41	3	95	10	3
42	5	90	10.95	2.83
43	4	76	16.73	2.4
44	8	88.88	17.64	2.39
45	4	88	17.89	2.8

Fig. 8. Results per video of the evaluation of the Arabic video summarization system.

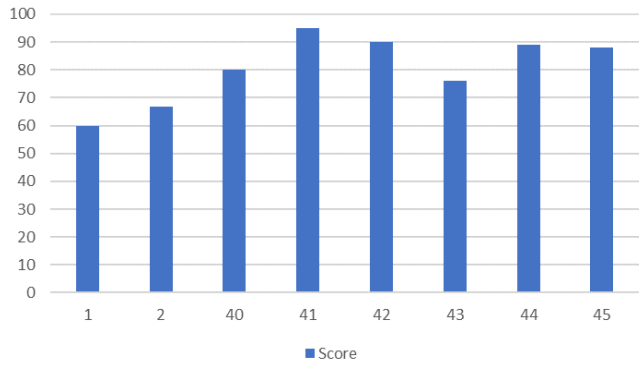

Fig. 9. Score per video of the evaluation of the Arabic video summarization system.

French Video Summarization System Evaluation Result.

Number of videos	Number of users	Number of answers	AVG score	AVG SD	GQ_AVG
10	6	11	64	2.82	1.93 (near good)

Fig. 10. Overall results of the evaluation of the French video summarization system.

Video ID	Number of users	AVG	SD	GQ_AVG
17	1	80	0	2
18	1	40	0	1
19	1	80	0	3
20	1	20	0	0.5
21	1	80	0	3
26	1	60	0	3.5
27	1	40	0	0.5
31	1	80	0	1.5
32	1	80	0	3
33	2	80	28.28	1.25

Fig. 11. Results per video of the evaluation of the French video summarization system.

10

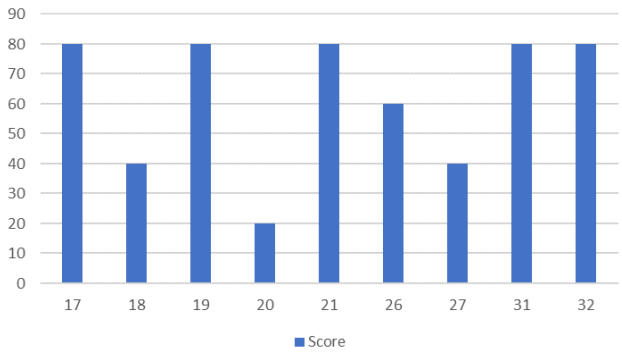

Fig. 12. Score per video of the evaluation of the French video summarization system.

4 Objective Evaluation

This section presents the evaluation results based on an objective (tag-based) evaluation procedure [2]. The aim of the objective evaluation is selecting a summarizing scenario based on the quality of different situations.

4.1 Input Data

This Subsection describes the input data that has been used. The video sequences are described separately, and the tag selection independently.

Video Sequences. The objective evaluation used the set of 27 video sequences, manually chosen to represent diversified content. From the source video sequences, 81 sequences were subsequently generated (for each source sequence, three series summarized, in scenarios 1, 3 and 4).

Tag Selection. For the tag-based evaluation procedure, the most critical data used were YouTube tags. We used YouTube tags for every selected video sequence. Tags used there, summaries the given newscast/report video sequence. Tickets are, depending on the particular video sequences, written in different languages. Here are some examples:

RT, Russia Today, FSA kicks out US special forces troops, FSA, Free Syrian Army, US special forces soldiers, withdraw, kicked out, Syria, war, troops, us-backed.

4.2 Evaluation Algorithm

Every video sequence used has its own set of tags. They are in different languages. The algorithm for each video sequence is:

1. Retrieve the audio track from the original video sequence.
2. Use an Automatic Speech Recognition (ASR) system engine [3]. It is a system created with two parts: an acoustic model and language model—the system bases on KALDI ASR toolkit. The acoustic model uses a Deep Neural Network (DNN), which has an input layer of 440 neurons, six hidden

layers of 2048 neurons each. The output layer has around 4000 neurons. ASR returns a set of words with timecodes. We use it to obtain a textual transcription of the original video sequence. Precise mathematical description and handling of the neural network approach has been provided in [1] and [3].

3. Retrieve tags. If a tab contains more than one word, split it. Create a set without duplicates.
4. Check which tags appear in the textual transcription of the original video sequence. Limit the set of cards to these tags that occur in the textual transcription of the original video sequence.
5. Create a summary of the original video sequence.
6. Retrieve an audio track from the recently created summary [7].
7. Use the ASR engine to obtain a textual transcription of the summarized video sequence.
8. Check which tags appear in the textual transcription of the summarized video sequence. Create a set of cards that occur in the textual transcription of the synthesized video sequence.
9. Check the tags that occur in the summarized and the original video sequence. Calculate statistics.

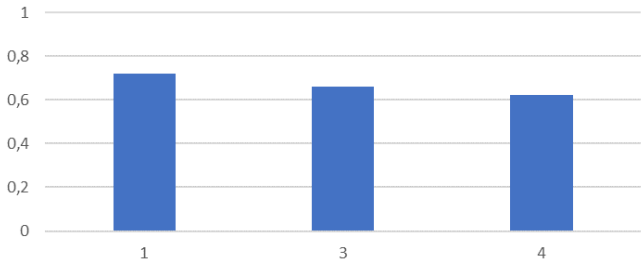

Fig. 13. Objective (tag-based) evaluation procedure results per scenario.

4.3 Experiment Results

This Subsection presents results for the objective, tag-based method. In Fig. 13, one can observe the relative number of tags in the summarized video sequences and the number of cards in the original video sequences – for all tested video sequences, grouped by scenarios.

As we can see, for most of the selected video sequences, the percentage of occurring tags in summaries is above 60. It means that they contain most of the content described in tags. It is a valuable check.

Regarding the winning scenario, it is not evident which one the responders indicated as the best one. All three scenarios gave more or less similar results; however, with a slightly visible indication of scene 1.

The difference between scenario one and the next scenario 3 is in practice five percentage points, which finally does allow to draw some slight conclusions.

The relatively high position of scenario 1 for objective evaluation is puzzling. While in the case of subjective assessment, the choice of a particular version of the summary is mostly a result of the respondent's taste, the objective evaluation, after all, includes the content of critical words in the review. Moreover, if so, it would be logical that the scenarios three and four should be the winning ones (which work just on the principle of focusing on compelling content), rather than ssituation1 (which only considers the visual aspect). In the meantime, it is not happening. So, the question is, wherefrom does this delicate advantage of Scenario 1 come?

Perhaps the answer stems from a particular feature of scenario 1, which does not appear in scenes 3 and 4. Now, in the scene, the first shot is rigorously included in the summary (in which, in the case of news and reports, it is not uncommon to find most valuable, crucial words). It naturally raises statistics in the case of objective evaluation. Meanwhile, scenario three and scenario four do not include such a rule, based solely on the analysis of the text. Consequently, the summary in inch consists first shot only in about 20% of cases. Perhaps, then, the quality of the summarizations generated in scenarios three and four could be increased using the rigid rule of incorporating the first shot, straight from scene.

5 Conclusion

Considering the performed evaluation about the different existing scenarios of the system and the assessment of the final system considering the situation that performed better, it is believed that it has been developed a complete evaluation procedure, which has provided exciting results for our policy, and that can be applied to other similar systems.

This deliverable reported two approached to the evaluation procedures, a subjective one and an objective one. As mentioned, the purpose of the described research was to try to narrow the space of possible scenarios. The actual evaluation did not indicate situations worse than others. Meanwhile, in the light of individual results obtained, the researches cannot unambiguously indicate one single scenario, recommended as the only one for further development. However, the researchers can state with certainty that the new development of scene 1, which has received many negative evaluations among professionals, should be discontinued. At the same time, as said, there is not a precise winner scenario. Still, considering the results obtained by the evaluations by the professionals, scene three has been considered as best scenario because of the obtained better results, even if the difference is little.

As a final subjective evaluation, scenario three has been used to generate the summaries of some videos, and different users have performed some tests to know how well the system functions. The obtained results, with an average above 70 out of 100, indicate that the performance of the system is excellent. It must be considered that the results of the tests, apart from the quality of the summary, depending on the understanding and individual capabilities of the evaluators.

It is considered that the objective of the project has been fulfilled and the value of the developed system is deemed to be high, as, it includes two different ways of evaluating a video summarization system evaluation, one that reflects the physical quality and other that reflects the real quality considered by the final users. It has been tested with a real project, getting promising results. So, the developed evaluation procedure can be used for the evaluation of any video summaries system, as a general use evaluation procedure has been established.

Acknowledgement

Research work funded by CHIST-ERA call 2014 (project AMIS under the topic Human Language Understanding: Grounding Language Learning).

References

1. Kamel Smaïli, Dominique Fohr, Carlos González-Gallardo, Michal Grega, Lucjan Janowski, Denis Jouvet, Arian Kozbial, David Langlois, Mikołaj Leszczuk, Odile Mella, Mohamed-Amine Menacer, Amaia Mendez, Elvis Linares Pontes, Eric Sanjuana, Juan-Manuel Torres-Moreno and Begoña Garcia-Zapirain. Summarizing videos into a target language: methodology, architectures and evaluation. Journal of Intelligent & Fuzzy Systems. 2019.
2. Kamel Smaïli, Dominique Fohr, Carlos-Emiliano González-Gallardo, Michał Grega, Lucjan Janowski, Denis Jouvet, Artur Komorowski, Arian Kozbial, David Langlois, Mikołaj Leszczuk, Odile Mella, Mohamed A. Menacer, Amaia Mendez, Elvis Linares Pontes, Eric Sanjuana, Damian Swist, Juan-Manuel Torres-Moreno, and Begoña Garcia-Zapirain. A first summarization system of a video in a target language. In Kazimierz Choros, Marek Kopel, Elzbieta˙ Kukla, and Andrzej Sieminski, editors, Multimedia and Network Information Systems, pages 77–88, Cham, 2019. Springer International Publishing.
3. Artur Komorowski, Lucjan Janowski, and Mikołaj Leszczuk. Evaluation of multimedia content summarization algorithms. In Kazimierz Choros, Marek Kopel, Elzbieta Kukla, and Andrzej Sieminski, editors, Multimedia and Network Information Systems, pages 424–433, Cham, 2019. Springer International Publishing.
4. Flora Amato, Aiello Castiglione, Vincenzo Moscato, Antonio Picadillo, Giancarlo Sperlì. Multimedia summarization using social media content. Multimedia Tools and Applications. 2018.
5. Interject Mani, David House, Gary Klein, Lynette Hirschman, Therese Firmin, Beth Sondheim. The TIPSTER SUMAC Text Summarization Evaluation. Proceedings of EACL '99.
6. Elena Lloret, Laura Plaza, Ahmet Aker. The challenging task of summary evaluation: an overview. Language Resources and Evaluation. 2018.
7. Denis Jouvet, David Langlois, Mohamed Amine Menacer, Dominique Fohr, Odile Mella, and Kamel Smaïli. About vocabulary adaptation for automatic speech recognition of video data. In ICNLSSP'2017 - International Conference on Natural Language, Signal and Speech Processing, pages 1–5, Casablanca, Morocco, December 2017.