Towards Automatic Captioning of University Lectures for French Students who are Deaf

Solène Evain, Benjamin Lecouteux, François Portet, Isabelle Estève, Marion Fabre

Towards Automatic Captioning of University Lectures for French Students who are Deaf

SOLÈNE EVAIN, BENJAMIN LECOUTEUX, and FRANÇOIS PORTET, Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, France
ISABELLE ESTEVE, Univ. Grenoble Alpes, LIDILEM, France
MARION FABRE, Univ. Lumière Lyon 2, ECP, France

Access to higher education of students who are deaf is below the national average. Recently, there has been a growing number of applications for the automatic transcription of speech, which claim to make everyday speech more accessible to people who are Deaf or Hard-of-Hearing. However, these systems require a good command of the written language, and a significant proportion of the deaf public has low literacy skills. Moreover, we have very little data on how these audiences actually deal with captions. In this paper, we describe the MANES project, whose long-term goal is to assess the usefulness of captioning for the accessibility of lectures by students who are deaf. We present the first technical results of a real-time system to make course captioning suitable for the target audience.

CCS Concepts: • **Human-centered computing** → **Accessibility technologies**.

Additional Key Words and Phrases: speech recognition, hearing loss, accessibility

ACM Reference Format:
Solène EVAIN, Benjamin LECOUTEUX, François PORTET, Isabelle ESTEVE, and Marion FABRE. 2018. Towards Automatic Captioning of University Lectures for French Students who are Deaf. In *Proceedings of Assets 2020: 22nd International ACM SIGACCESS Conference on Computers and Accessibility (Assets 2020).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Nowadays, only 5% of people who are deaf access higher education [3]. Promoting accessibility to higher education for students who are deaf is a priority of the French government since the 2005 law [1] and is one of the key measures of the 2010-2012 program in favor for people who are deaf or hard-of-hearing [7]. Despite the legal obligation to adjust training programs to students with disabilities, no plan for inclusive pedagogy has been implemented yet for students who are deaf.

Interpreters in French signed language (LSF), coders in cued speech (LPC), transcribers with velotypes are among the many solutions for students who are deaf to access university courses. However, this requires financial and human resources which some universities cannot afford. Automatic captioning emerged recently thanks to high-performance automatic speech recognition (ASR) systems and might be seen as an accessibility solution, either on its own or as a complementary source to LSF [12]. However, if there have been much research for English language [6] [11], we only found one study exploring the use of captioning obtained via ASR for French lectures, using 're-voicing' and collaborative editing [5]. As for already existing ASR system such as RogerVoice or Google, they state that such good

Manuscript submitted to ACM

ASR systems permit a larger accessibility to oral communication in everyday life for people with hearing disorders. Yet, the users are supposed to have a good mastery of reading and a large part of the deaf community has low literacy level. So far, there have been little research about how to format transcriptions for this particular audience (e.g. highlighting keywords [4], [9]). Furthermore, despite ASR systems good performance, decoding errors are still problematic for captions comprehension [8], especially for students who are deaf who "do not read by 'sounding the words out'". Finally, big companies' systems are not configurable enough and are ethically questionable for a use in an academic context. Our objective is then to design, develop and study a real-time captioning system for students who are deaf to improve accessibility to lectures and self-appropriation of knowledge by providing adapted captions and lowering transcription errors due, for instance, to unknown vocabulary (OOV words).

2 PROJECT

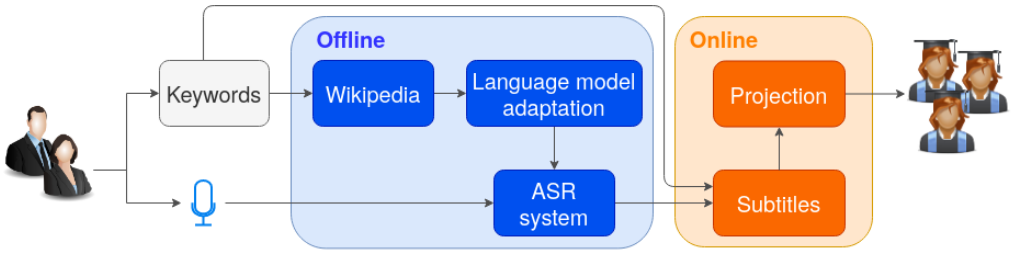

Fig. 1. MANES captioning system overall presentation

Figure 1 is an overall view of the captioning system. Before class the teachers produce a list of keywords that represent important concepts or names in their course. Those keywords are used for collecting Wikipedia articles or from given lecture documents that are topic-related. The retrieved documents are used for language model adaptation of the ASR system. The aim is to add words to the vocabulary and to extract patterns that are relevant for the course (for e.g. 'bubble sort algorithm' for an algorithmic course). Those patterns are then assigned a higher probability to be recognized by the system. During class, the real-time ASR system is used to generate adapted captions. The keywords used to create the Wikipedia corpus are used to highlight important concepts in the captions.

3 METHOD

3.1 Development of an automatic captioning system

An automatic captioning system was developed thanks to the ASR Kaldi toolkit [10]. The first added value of our system is the spotting of keywords to make the understanding and extraction of lexical, semantic and syntactic clues easier and help students who are deaf with the note-taking. The second one is the adaptation of the ASR system (Figure 2). Indeed, ASR systems can only transcribe words that are in their lexicon. Creating a thematic corpus in relation to the course permits to add new vocabulary to the lexicon and increase the probability of n-grams that are domain-specific. This helps improving recognition results, leading to better understanding of captions.

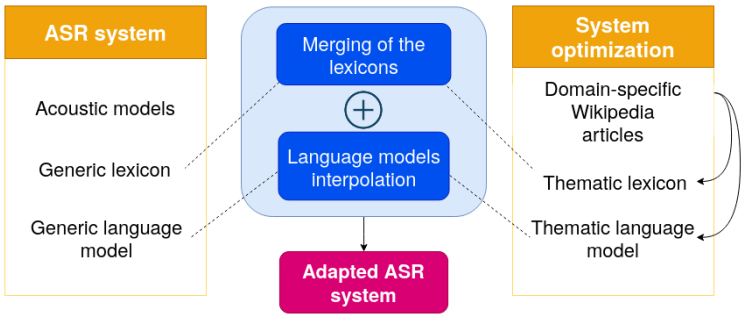

Fig. 2. ASR system adaptation

3.2 Data

A 3-gram *generic language model* was trained from several French corpora: EUbookshop, TED2013, Wit3, GlobalVoices, Gigaword, Europarl-v7, MultiUN, Opencaptions2016, DGT, News Commentary, News WMT, LeMonde, Trames, Wikipedia. The acoustic model was trained from ESLO2 (126 hours of spontaneous speech) and ESTER2 (about 100 hours of radio broadcast: news, talks) corpora.

We also collected the following corpora for course-related language model training and decoding tests:

- transcribed recordings of university lectures (algorithmic, automatic translation, information retrieval, speech-to-speech translation): spontaneous noisy speech recorded with fixed microphone (about 4h19)[1].
- Wikipedia articles collected thanks to course-related keywords (6 articles, 10167 words, 235 different words).

4 FIRST RESULTS

4.1 A captioning system for the deaf

A graphical interface was developed. It permits to load an complete a list of keywords or record and save a new list. Those keywords will be highlighted in the caption. The teacher launches the captioning via the interface. At the end of the class, he/she saves the '.wav' recording of the course and the captions in a '.vtt' format. He/she has the possibility to correct the captions before sharing them with the students. The example of a '.vtt' file on the right presents the obtained captions for 'Hello everyone my name is Eric, I am a teacher in computer science at the university and I work at the computer laboratory of Grenoble'. Each contains 2 lines max, 40 characters per line on average, highlighted keywords. Filler words are removed. The transcriptions are generated with temporal and spatial constraints: last at least 1 second on screen, check for update every 40 milliseconds, cut >40 characters sentences in two without cutting a word and cut >90 characters sentences into two captions. We set a 10 seconds limit of recording for one 'sentence' to Kaldi online command.

```
WEBVTT

1
00:00:00.000 --> 00:00:05.010
hello everyone my name is éric i am a
teacher in <u>computer science</u> at the university

2
00:00:05.010 --> 00:00:10.080
and i work at the computer laboratory of
grenoble
```

[1]Transcriptions have been made following the transcription conventions of the VALIBEL corpus [2]

4.2 Adaptation results

The adaptation of the language model (LM) was measured using perplexity (ppl). It calculates how much a language model is able to predict words of a text that was not used for its training (the lower the ppl, the better). The evaluation corpus is a university lecture about algorithmic. A 248 perplexity is obtained with the generic LM only. As you can see in Table 1 the interpolation of one or two topic-related LMs with the generic one is beneficial only if the generic LM interpolation coefficient is lower than the other(s). The best result (147.502 ppl) is obtained when B and C LMs are separated and have different coefficients. The B LM is composed of transcribed courses from the same teacher as the evaluation corpus. This shows that not only thematic data set is important but also user speech style representation.

A: generic LM, B: university courses, C: Wikipedia articles, B/C:University courses + Wikipedia articles in one LM

LM	Interpolation coefficients (sum to 1)	Ppl	LM weight	WER
A	x	247.996	1	40.09%
A + B/C	B/C: 0.1	275.147	1	51.32%
A + B/C	A: 0.26, B/C: 0.74	208.957	1	45.16%
A + B/C	A: 0.26, B/ C: 0.74	208.957	8	39.38 %
A + B + C	A: 0.08, B: 0.43, C: 0.49	147.502	8	33.29%

Table 1. Perplexity measures for different Language Models and ASR performance.

4.3 Recognition results

The Word Error Rate (WER) was used for decoding performance evaluation (the lower, the better). As you can see in Table 1, the generic LM gives a 40.09% WER. Second and third LMs worsen the results. Increasing the LM weight for the third one gives a slightly better WER than reference (generic LM): -0.71 point. Finally, the best WER (33.29%) is achieved with the LM configuration that had the best perplexity, with a LM weight set to 8. We classified the substitution errors happening more than once in 4 classes (see Table 2).

OOV words	Grammatical err.	Homophonic err.	Others	Total subst.
3.41%	19.97%	35.14%	41.49%	646

Table 2. Most frequent transcription errors.

We called 'homophonic errors' substituted words that are homophonic or paronymic. 'Others' contains errors that may be due to errors in manual transcription, errors in vocabulary or uncategorisable errors (e.g. seven –> one) that may appear when alignment fails. It turns out that plenty of errors are due to homophony or paronymy (35.14%). A 33% WER is not good enough for tests but it could be lowered with more specific LM (ideal case with transcriptions of the test set used for LM interpolation gives a 4.61% WER).

5 FUTURE WORK :

Decoding could be improved with higher quality and less noisy recordings, so more datasets are needed. Indeed, in our test set, the teacher had a fixed mic rather than a headset. So signal is degraded when he moves and students are heard chitchatting. Using the generated captions for improving the language model for LM "speaker adaptation" will give better results as we saw in Table 1 with the C LM. We could also imagine post-class collaborative editing by hearing students to correct the captions. Future tests will be made to predict [4] and evaluate the impact of captions on students who are deaf during a lecture. Those tests will also give us clues for an optimized formatting of our captions.

REFERENCES

[1] [n.d.]. Loi n° 2005-102 du 11 février 2005 pour l'égalité des droits et des chances, la participation et la citoyenneté des personnes handicapées.

[2] Sylviane Bachy, Anne Dister, Michel Francard, Geneviève Geron, Vincent Giroul, Philippe Hambye, Anne-Catherine Simon, and Régine Wilmet. 2007. Conventions de transcription régissant les corpus de la banque de données VALIBEL. https://dial.uclouvain.be/pr/boreal/object/boreal:165551 Number: UCL - Université Catholique de Louvain.

[3] Laurence Haeusler. 2014. Étude quantitative sur le handicap auditif à partir de l'enquête "Handicap-santé".

[4] Sushant Kafle, Peter Yeung, and Matt Huenerfauth. 2019. Evaluating the Benefit of Highlighting Key Words in Captions for People who are Deaf or Hard of Hearing. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)*. Association for Computing Machinery, Pittsburgh, PA, USA, 43–55. https://doi.org/10.1145/3308561.3353781

[5] Benoît Lathière and Dominique Archambault. 2014. Improving Accessibility of Lectures for Deaf and Hard-of-Hearing Students Using a Speech Recognition System and a Real-Time Collaborative Editor. In *Computers Helping People with Special Needs (Lecture Notes in Computer Science)*, Klaus Miesenberger, Deborah Fels, Dominique Archambault, Petr Peňáz, and Wolfgang Zagler (Eds.). Springer International Publishing, Cham, 490–497. https://doi.org/10.1007/978-3-319-08599-9_73

[6] William McKee and David Harrison. 2008. Producing Sub-titles Lecture Recordings for the Deaf using Low-cost Technology. Liverpool Hope University, 152–156.

[7] de la famille de la solidarité et de la ville Ministère du travail, des relations sociales and Secrétariat d'état chargé de la famille et de la solidarité. 2010. Plan en faveur des personnes sourdes ou malentendantes - 1....

[8] Becky Parton. 2016. Video Captions for Online Courses: Do YouTube's Auto-generated Captions Meet Deaf Students' Needs? *Journal of Open, Flexible, and Distance Learning* 20, 1 (Aug. 2016), 8–18. https://www.learntechlib.org/p/174235/ Publisher: Distance Education Association of New Zealand.

[9] Agnès Piquard-Kipffer, Odile Mella, Jérémy Miranda, Denis Jouvet, and Luiza Orosanu. 2015. Qualitative investigation of the display of speech recognition results for communication with deaf people. In *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*. Association for Computational Linguistics, Dresden, Germany, 36–41. https://doi.org/10.18653/v1/W15-5107

[10] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlıcek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi Speech Recognition Toolkit. Hawaii, US, 4.

[11] Rohit Ranchal, Teresa Taber-Doughty, Yiren Guo, Keith Bain, Heather Martin, J. Paul Robinson, and Bradley S. Duerstock. 2013. Using speech recognition for real-time captioning and lecture transcription in the classroom. *IEEE Transactions on Learning Technologies* 6, 4 (Oct. 2013), 299–311. https://doi.org/10.1109/TLT.2013.21 Conference Name: IEEE Transactions on Learning Technologies.

[12] Joong-O Yoon and Minjeong Kim. 2011. The Effects of Captions on Deaf Students' Content Comprehension, Cognitive Load, and Motivation in Online Learning. *American Annals of the Deaf* 156, 3 (2011), 283–289. https://www.jstor.org/stable/26235157 Publisher: Gallaudet University Press.