Building and enhancement of an ASR system for emergency medical settings: towards a better accessibility for allophone and disabled patients

LUCÍA ORMAECHEA GRIJALBA, JOHANNA GERLACH, DIDIER SCHWAB, PIERRETTE BOUILLON, and BENJAMIN LECOUTEUX, Laboratoire d'Informatique de Grenoble and Université de Genève

In this article we aim to present the adaptation of an automatic speech recognition system for specific and robust applications related to the fields of medicine and disability. It constitutes the first step towards the building of an open source system designed to automatically translate speech into pictograms for allophone speakers or people having cognitive disorders in emergency medical settings. Due to the criticality of an adequate speech recognition in such contexts, we decided to rely on formal grammars representing the natural language used. Some preliminary experiments are displayed in order to evaluate its use in real-life situations.

Additional Key Words and Phrases: Automatic Speech Recognition, Pictogram Transcription, BabelDr

ACM Reference Format:
Lucía Ormaechea Grijalba, Johanna Gerlach, Didier Schwab, Pierrette Bouillon, and Benjamin Lecouteux. 2020. Building and enhancement of an ASR system for emergency medical settings: towards a better accessibility for allophone and disabled patients. 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.XXXX/XXXXXX.XXXXXXX

1 INTRODUCTION

In present-day health services, language barriers are a problem of undeniable seriousness, since the difficulty or impossibility of a good interaction between a person being treated and a physician who do not share a common language can lead to fatal repercussions [8]. This particular issue is further accentuated in the case where a treated person has a cognitive disability, in which communicating with the specialist may be subject to other limitations. It is precisely for this reason that reliable mechanisms for specialized medical translation and interpretation are significant, in order to favor an adequate and effective communication between the actors above mentioned.

It is in this light that the BabelDr project was born, with the aim of guaranteeing effective assistance in multilingual medical emergency contexts and eliminating as far as possible the linguistic limits between the medical service and the patient. At runtime, the speaker's voice is recognized with the help of an Automatic Speech Recognition (ASR) system currently provided by *Nuance*. The resulting spoken utterance is then linked to a canonical form, which stands as a pivot for translation and as a backtranslation to verify recognition [4], and is considered as the input to be translated and read aloud into the target language.

The project currently offers translations into languages such as Arabic, Spanish or Farsi, but also works with minority languages like Tigrinya or Albanian. For the purpose of enhancing its accessibility, BabelDr is working on the inclusion of Norman Switzerland deaf sign language and envisions to move towards a system that would likewise translate

Manuscript submitted to ACM

utterances into pictograms [11, 17, 20]. It would not only be useful as a means of patient-doctor communication for allophone speakers (namely non-French speakers), but also for people having a cognitive disability.

The present paper explores the first step towards the development of a BabelDr open-source version, which involves the building and enhancement of an ASR system that uses the Kaldi software toolkit and employs language models based on formal grammars built by translation specialists. The remainder of this article is as follows. We will first contextualize the BabelDr project in Section 2 and then describe our methodology in Section 3. We will then present the evaluation corpora being used and the experiments we carried out in Section 4. Finally, we will analyze the obtained WER results in Section 5, followed by the conclusions on this work.

2 THE CASE OF BABELDR

Multiple translation platforms are used within medical settings. Telephone interpreting services are among the most popular, but they are expensive [4] and often not available for all languages for which translation is required [19]. For their part, free online automatic translation platforms such as Google Translate are not adapted to a language for specific purposes and do not guarantee sufficient data protection [5] nor secure cloud storage.

This is why specialized translation tools such as BabelDr have arisen, so as to provide a reliable and intuitive translation tool for emergency medical settings. With the help of an automatic voice recognition system, BabelDr facilitates two-way patient-physician interaction [3]. Although based on a pre-established set of sentences translated by humans, the main difference to regular phraselators (such as MediBabble or Canopy) is that doctors can find the desired sentence by speaking to the system orally and using a wide range of paraphrases and stylistic variations [5], and thus not being obliged to search them manually in a predefined list.

In order to implement such a translation tool, an appropriately designed ASR system is required. More precisely, this implies adapting it to properly recognize a language for specific purposes such as medical discourse which, from a lexical perspective, is characterized by its borrowed terminology from Greek and Latin and the widepread use of neologisms. And syntactically speaking, its particular phraseology is also noticeable in the way that the specialist refers to the symptoms or determines a diagnosis [16]. Currently, BabelDr ASR system is blackbox-based and is provided by Nuance. The software is hosted at the hospital, but the medical personnel do not have control over it to improve or modify it. Our work aims to promote independence for ASR through the use of open tools.

3 METHODOLOGY

In order to build an ASR system for medical purposes[1], we chose to use the open-source Kaldi toolkit [13]. Kaldi uses Finite State Transducers (FSTs) as a framework to build language models (LMs), whose very aim is to predict probable word sequences during decoding [1]. Work presented in [9] has provided a basis for this study.

Since the Finite State Grammars (FSG) created by translation specialists were written in a source format according to the Regulus Lite formalism [15], they could not be directly used by our system. This thus made necessary a conversion into FST. To do so, we first proceeded to a normalization. This way, we distinguished the main language model from the sub-language models, the former being composed of a set of utterances that model medical discourse by considering different paraphrases, and the latter being embodied by two types of word classes. *TrLex* classes represent lexical paradigms and *TrPhrase* phrasal patterns; they are both identified by non-terminal symbols ($) and can be integrated into the main word language model. An example of the source file format is the following:

- Utterance : Utterance $avez_vous (mal | douleur) (quelque part | à un endroit) EndUtterance
- Phrasal patterns : TrPhrase $avez_vous (avez-vous | vous avez) EndTrPhrase

[1]All the tools to reproduce experiments will be made available in: https://github.com/lormaechea/Grammar-Tools-ASR

The steps below need to be followed in order to render the grammar usable in Kaldi:

(1) Convert every single pattern into a FST and compiled them separately using `fstcompile`.
(2) All the compiled FSTs belonging to the main language model are unified into a single compiled file, whereas for the sub-language models, a differentiated unification of each word class is operated, using `fstunion`.
(3) When integrating the word classes into the main grammar, recursively replace all the existent non-terminal symbols (starting with $) into the main language model with the according terminals by `fstreplace`.

On this basis, there is a resulting `G.fst` that contains the main grammar including the word classes but it is not operational in Kaldi yet. The next phases need to be followed:

(1) Remove all epsilon transitions (`<eps>:<eps>`, no input and output symbols) with `fstrmepsilon`.
(2) Determinize the resulting FST with `fstdeterminize`.
(3) Minimize it with `fstminimizeencoded`, one of the extra Kaldi OpenFST executables. If `fstminimize` was used instead, it would automatically push the labels to the start where possible, which could result into shifting the input and output label combination in the FST.
(4) Sort the arcs in the resulting FST per state with `fstarcsort`.
(5) After this procedure, the generated model can be directly used in Kaldi.

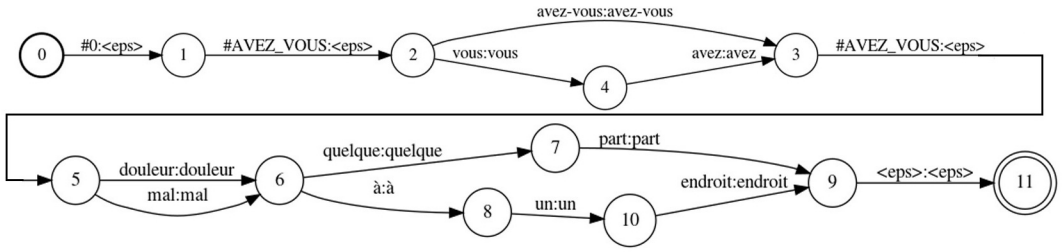

Fig. 1. This image shows the resulting FST for the preceding Utterance and TrLex example.

4 AUTOMATIC SPEECH RECOGNITION EXPERIMENTS AND CORPORA

Our ASR system is hybrid HMM-DNN based and is trained with lattice-free MMI [14], using the Kaldi toolkit [13]. The acoustic model (AM) topology consists of a Time Delay Neural Network (TDNN) followed by a stack of 16 factorized TDNNs [12]. The acoustic feature vector is a concatenation of 40-dimensional MFCCs without cepstral truncation (MFCC-40) and 100-dimensional i-vectors for speaker adaptation [6]. Audio samples were randomly perturbed in speed and amplitude during the training process [10]. We trained the system following the *tedlium* recipe[2]. The TDNN layers have a hidden dimension of 1536 with a linear bottleneck dimension of 160 in the factorized layers. The i-vector extractor is trained on all acoustic data (speech perturbed + speech) using a 10s window. The acoustic training data includes 500 hours of French speech: ESTER [7], COMMON-VOICE [2] and ESLO [18]. It should be noted that the recognition system operates in real time (online configuration) and produces hypotheses on the fly.

4.1 A dedicated audio corpus for evaluation

As seen below, the evaluation corpora are heterogeneous. For the dev set we use *exp1Stud* and *exp1Doct*, both including data collected with an older BabelDr version (Flash client), *exp1GT* containing data collected when using Google Translate (video recording of complete sessions, where the sentences pronounced by the doctors are extracted) and the

[2]https://github.com/kaldi-asr/kaldi/tree/master/egs/tedlium

ellipsis corpus containing artificial data coming from a similar online platform to the present BabelDr client. The test set, *exp2*, contains data that was gathered with the current version of BabelDr (JavaScript client).

	dev				*test*
	ellipsis	**exp1Doct**	**exp1Stud**	**exp1GT**	**exp2**
# Words	10266	1755	3107	1961	5571
# Utterances	1642	334	573	321	928
# Speakers	5	6	4	9	12
# Speech	1h13	0h20	0h26	0h16	0h57
WER cLM	23.02 [38.27]	28.69 [30.25]	31.03 [39.39]	29.65 [40.28]	**28.97** [41.34]
WER uLM	13.00 [13.60]	22.51 [15.06]	20.83 [11.46]	16.43 [13.57]	**18.89** [13.23]
Overall WER	**cLM**:	**22.70 [37.53]**	**uLM:**	**16.30 [13.40]**	–

Table 1. Results of our ASR system on our different datasets using unconstrained (uLM) and constrained LMs (cLM). The overall WER is calculated from all the data and weighted by their proportions. The numbers in brackets correspond to the Nuance system results.

4.2 Experiments using: constrained vs unconstrained grammars

In first experiments, an ASR system strictly based on the compiled FSG was tested. The results reveal that spoken utterances belonging to G were far more correctly recognized that the ones differing from it. This explains why sometimes the transcription deviates from what was pronounced: the system is misled because it intends to respond exactly to the FSG possible outputs. This is why we tried to relax the constraints related to G by creating a n-gram LM based on the strings generated by it. Table 1 presents the results of the two systems.

5 RESULTS ANALYSIS AND DISCUSSION

We observe that the WER obtained with the constrained grammar is higher. This is explained by the fact that as soon as the user moves away from the grammar, the system cannot adapt and outputs the closest path in the grammar. However, in almost all cases the ASR system outperforms the Nuance system. For the unconstrained version, we observe a much lower error rate. However, the outputs seem more difficult to exploit in a robust approach as a translation must match a sentence generated by the grammar. In such cases, we see that Nuance's system obtains a better WER. This is probably due to the fact that our unconstrained system is only trained on grammar data, whereas the Nuance system is more generic: future work will focus on the joint use of a constrained LM with a more open model so as to follow the optimal strategy according to the situation.

6 CONCLUSION AND PERSPECTIVES

This work shows that it is possible to use an open source ASR system as part of a Speech-to-Text application constrained by a grammar. The BabelDr[3] system currently uses an ASR system based on Nuance's blackbox tools; the results we get are similar from an end-user perspective. This means that the proposed approach will soon be ready for deployment in production. The next steps will focus on integrating the possibility of translating directly from speech to pictograms for accessibility purposes. In the longer term, we would like to extend this type of approach to environments outside the medical setting, such as family or school.

[3]**Acknowledgments:** We would like to thank the University Hospitals of Geneva and all the doctors who contributed to the recording of the corpus used for our tests.

REFERENCES

[1] Martine Adda-Decker and Lori Lamel. 2000. The use of lexica in Automatic Speech Recognition. In *Lexicon Development for Speech and Language Processing*, Nancy Ide, Jean Véronis, Frank Van Eynde, and Dafydd Gibbon (Eds.). Springer Netherlands, Dordrecht, 235–266. http://link.springer.com/10.1007/978-94-010-9458-0_8

[2] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2019. Common Voice: A Massively-Multilingual Speech Corpus. arXiv:1912.06670 [cs.CL]

[3] Pierrette Bouillon, Glenn Flores, Marianne Starlander, Nikos Chatzichrisafis, Marianne Santaholma, Nikos Tsourakis, Manny Rayner, and Beth Ann Hockey. 2007. A bidirectional grammar-based medical speech translator. In *Proceedings of the Workshop on Grammar-Based Approaches to Spoken Language Processing*. Prague, 42–48. http://portal.acm.org/citation.cfm?doid=1626333.1626341

[4] Pierrette Bouillon, Johanna Gerlach, Hervé Spechbach, Nikos Tsourakis, and Sonia Halimi. 2017. BabelDr vs Google Translate: a user study at Geneva University Hospitals (HUG). In *Proceedings of the 20th Annual Conference of the European Association for Machine Translation*. Prague. https://archive-ouverte.unige.ch/unige:94511/

[5] Valérie Boujon, Pierrette Bouillon, Hervé Spechbach, Johanna Gerlach, and Irene Strasly. 2018. Can speech-enabled phraselators improve healthcare accessibility? A case study comparing BabelDr with MediBabble for anamnesis in emergency settings. In *Proceedings of the 1st Swiss Conference on Barrier-free Communication*. Winterthur, 50–65. https://archive-ouverte.unige.ch/unige:105852/

[6] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. 2010. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing* (2010).

[7] S. Galliano, E. Geoffrois, G. Gravier, J. f. Bonastre, D. Mostefa, and K. Choukri. 2006. Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In *In Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC 2006*. 315–320.

[8] Karen Hacker, Maria Elise Anies, Barbara Folb, and Leah Zallman. 2015. Barriers to health care for undocumented immigrants: a literature review. *Risk Management and Healthcare Policy* (2015), 175–183.

[9] Axel Horndasch, Caroline Kaufhold, and Elmar Nöth. 2016. How to add word classes to the kaldi speech recognition toolkit. *In International Conference on Text, Speech, and Dialogue* (2016), 486–494.

[10] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. https://www.danielpovey.com/files/2015_interspeech_augmentation.pdf

[11] Magali Norré, Pierrette Bouillon, Johanna Gerlach, and Hervé Spechbach. 2020. Evaluating the comprehension of Arasaac and Sclera pictographs for the BabelDr patient response interface. *Barrier Free Conference, BFC* (2020).

[12] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur. 2018. Semi-orthogonal low-rank matrix factorization for deep neural networks.

[13] 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.

[14] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. 2016. Purely sequence-trained neural networks for ASR based on lattice-free MMI.

[15] Manny Rayner, Alejandro Armando, Pierrette Bouillon, Sarah Ebling, Johanna Gerlach, Sonia Halimi, Irene Strasly, and Nikos Tsourakis. 2016. Helping domain experts build phrasal speech translation systems. In *Future and Emergent Trends in Language Technology*, José F. Quesada, Francisco-Jesús Martín Mateos, and Teresa Lopez-Soto (Eds.). Vol. 9577. Springer International Publishing, Cham, 41–52. https://doi.org/10.1007/978-3-319-33500-1_4 Series Title: Lecture Notes in Computer Science.

[16] Maurice Rouleau. 2007. La langue médicale : une langue de spécialité à emprunter le temps d'une traduction. *TTR : traduction, terminologie, rédaction* 8, 2 (2007), 29–49. http://id.erudit.org/iderudit/037216ar

[17] Didier Schwab, Pauline Trial, Céline Vaschalde, Loïc Vial, and Benjamin Lecouteux. 2020. Providing semantic knowledge to a set of pictograms for people with disabilities: a set of links between WordNet and Arasaac: Arasaac-WN. *LREC* (2020).

[18] Noëlle Serpollet, Gabriel Bergounioux, Annie Chesneau, and Richard Walter. 2007. A large reference corpus for spoken French: ESLO 1 and 2 and its variations. (01 2007).

[19] Hervé Spechbach, Johanna Gerlach, Sanae Mazouri Karker, Nikos Tsourakis, Christophe Combescure, and Pierrette Bouillon. 2019. A speech-enabled fixed-phrase translator for emergency settings: crossover study. *JMIR Medical Informatics* 7, 2 (2019). http://medinform.jmir.org/2019/2/e13167/

[20] Céline Vaschalde, Pauline Trial, Emmanuelle Esperança-Rodier, Didier Schwab, and Benjamin Lecouteux. 2018. Automatic pictogram generation from speech to help the implementation of a mediated communication. In *Conference on Barrier-free Communication*. Geneva, Switzerland. https://hal.archives-ouvertes.fr/hal-01880744