MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231048

Development of an ASR System for Medical Conversations

Alejandro RENATO[a,1], Daniel LUNA[a] and Sonia Benítez[a]
[a]*Department of Helth Informatics, Hospital Italiano de Buenos Aires*
ORCiD ID: Alejandro RENATO https://orcid.org/0000-0001-6647-1268, Daniel LUNA https://orcid.org/0000-0001-5468-9761, Sonia BENITEZ https://orcid.org/0000-0001-6648-1984

Abstract. In this work we document the development of an ASR system for the transcription of conversations between patient and doctor and we will point out the critical aspects of the domain. The system was trained with an acoustic base of spontaneous speech that has a domain language model and a supervised phonetic dictionary. Its performance was compared with two systems: a) NeMo End-to-End Conformers in Spanish and b) Google API ASR[2] Cloud. The evaluation was carried out on a set of 208 teleconsultations recorded during the year 2020. The WER (Word Error Rate) was evaluated in ASR, and Recall and F1 for recognized medical entities. In conclusion, the developed system performed better, reaching 72.5% accuracy in the domain of teleconsultations and an F1 for entity recognition of 0.80.

Keywords. Automatic speech recognition, artificial intelligence, human-computer interaction, natural language processing, telemedicine

1. Introduction

The transcription of the conversation between doctor and patient constitutes valuable information for health institutions. The entities contained in this context, such as symptoms, exams, medications, vital signs, can be used both for the automatic construction of the medical evolution [1], for statistical purposes or as another record of the electronic medical record. In this sense, Kodish-Wachs [2] refers to two articles [3] and [4], where the recordings of the medical interviews and the resulting evolutions are compared and "significant omissions" were found and an accuracy of 71%-73% for the diagnostic category.

For example, the HCEM [5] system developed for the medical report dictation has a WER (Word Error Rate) of 6%, while the same system applied to the doctor-patient conversations of the present study drops drastically to the level of 35%. A recent study found that the WER of simulated medical conversations with a commercial Automatic Speech Recognition (ASR) engines was between 65% and 34% [2]. The problem is that these systems are trained for controlled dictation situations and with a high frequency of use of vocabulary. In a recent work [1], a patient-doctor transcription model trained with

[1] Corresponding Author: Alejandro Carlos Renato, Department of Health Informatics, Hospital Italiano de Buenos Aires, Faculty of Medicine (UBA, University of Buenos Aires), email: alejandro.renato@hospitalitaliano.org.ar/ arenato@fmed.uba.ar.

[2] ASR: Automatic Speech Recognition

thousands of hours of parallel audio and medical documentation is presented and reports the best performance, with a WER of 18.3%. However, data is not always available to developers, due to data privacy protocols.

In addition to the difficulties indicated in spontaneous speech, such as false starts, filled pauses, and unfinished sentences [6-8], there are other factors involved in poor performance, such as the speaker distance from the microphone or the low computer recording quality, voice overlapping, and ambient noise [9,10]. In addition, the problems produced during diarization [11] which segments the recording into speech turns per speaker should be pointed out.

At the moment, automatic recognition systems can present a variety of architectures in which End-to-End systems (E2E onwards) predominate, in which during training the system learns to predict the letter or fragments of words such as syllables or morphemes, from the audio. In this sense, they do not require a specific module to model pronunciation and language models, although they can be added in decoding. This paper presents the development of a hybrid system (HMM/TDNN) that offers the possibility of using audios that don't belong to the domain but show characteristics of spontaneous speech and hospital population dialectal varieties. The language model was adapted to the medical domain and the pronunciations of the phonetic dictionary were manually supervised.

To evaluate the performance of the system, a set of 208 teleconsultations recorded during year 2020 was used and compared with a NeMo Conformer system for Spanish[3] (E2E) [12] and the Google ASR system in Spanish[4]. The evaluation was measured in WER for the ASR system, in Recall and F1 for recognized medical entities.

2. Methods

This research was executed from 2019 to 2021. The overall system performance depends on the automatic recognition system as a critical element [13], and that's the reason why we focus our attention in acoustic training and language model building.

2.1. Acoustic Training

A large corpus (42,000 hours) representative of the phonetic variation and dialects of Argentina was collected. Public domain audios (33,000 hours) were used with preference for spontaneous speech. In addition, audios from the medical domain were collected, as well as presentations by the Argentine Medical Societies (7,000 hours) and others from the Italian Hospital of Buenos Aires (2,000 hours). Then the audios were subjected to a cycle of various tasks such as forced alignment [1] of the transcripts generated in various decoding processes, and segments with low signal-noise level were evaluated. As a consequence, 6,000 hours were selected, which met the criteria of balance and confidence in the transcription.

The ASR system was trained using the factored TDNN model [14] using the Kaldi toolkit. The choice of the TDNN system responded to several reasons: a) the TDNN models have the advantage of being able to model acoustic and language models independently, extracted from different domains, b) the phones acoustic models can

[3] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_es_conformer_ctc_large

[4] https://cloud.google.com/speech-to-text?hl=es

easily incorporate phonetic, linguistic and contextual knowledge structured by using a lexicon and c) it does not require expensive infrastructure and has acceptable performance in real time.

2.2. Language Model

The language model was built to improve the adaptation to the domain of doctor-patient conversation, in which colloquial and interactive language is combined with medical vocabulary. A corpus of medical reports made at the Hospital Italiano of Buenos Aires between 2017 and 2021 (144 million sentences and 91,702 words vocabulary), duly anonymized (patients ID and doctors and patients names) was used. Pronunciations of recognized Spanish words were automatically generated by training the Phonetisaurus system, and foreign words, proper names, and abbreviations were corrected by hand. The rest of the corpus was collected from various sources (Spanish TED Talks, TV and Radio medicine programs and interviews, medical internet forums, parliamentary speeches and oral and public trials), with a prevalence of spontaneous speech transcripts. The search criteria were the high occurrence of first and second person pronouns and verbs, which reflected speech in a verbal interaction, which is not represented in medical reports. It is made up of 20 million sentences and 177 million words. The dictionary in total has 388,649 words. The model is generic: the vocabulary and the selected corpus was independent of the evaluation text, to avoid bias in the results. With the corpus, a quadrigram model was built by interpolating 20 models with 4grams using modified Kneser-Ney smoothing.

3. Results

The test set has 208 teleconsultations carried out during 2020 that were recorded online on video from Jitsi Meeting System and then the audio was extracted in OGG Vorbis format, mono, at 48Khz sampling rate. The objective of the recordings was to evaluate the system performance and the quality of video and audio transmission. The teleconsultations lasted 16 hours 35 minutes of audio, and the participants were: 12 doctors and 202 patients. Audio quality was poor, with abundant noise and voices overlapping. When more than a quarter of the emission presented overlapping voices or with high intensity noise, the wave was discarded. The microphones used in many cases are not adequate for ASR, such as the microphone present in cameras or notebooks. The "BBC Speech Segmenter" system [15] was used for the diarization of teleconsultations. The test data set was not used in the development of the system, nor were similar data.

Table 1. ASR WER and Accuracy, and Entity Recognition Precision Recall and F1.

Model	Kaldi TDNN	NeMo Conformer Spanish	Google API Spanish
Accuracy/WER	72.5 % / 27.5	64 % / 36	39 % / 61
Precision (Entities)	0.98	0.97	0.86
Recall (Entities)	0.87	0.78	0.42
F1 (Entities)	0.80	0.42	0.28

WER and Accuracy was carried out with the software sctk[5]. Medical entities were evaluated as correct when the ASR transcribed them appropriately. No entity recognition system was evaluated here.

4. Discussion

The Kaldi TDNN system obtained better results compared to the other two systems and achieved 72.5% of accuracy and a F1 = 0.80 in medical entities recognition. Although the WER continues to be high, the errors are concentrated in connectors (articles, prepositions) or in morphological variants (plural vs. singular), while the lexemes that are important for the recognition of entities and contents are preserved. Its comparative advantage is that it can be trained in acoustic terms in similar conditions to the hospital environment, with a population that has the dialectal variations of the patients, allowing control of pronunciations and language models.

The E2E restricts the training vocabulary, although it is possible to improve its performance by coupling a language model. As a consequence, it had many errors as creating non-existent words that have pronunciations similar to the right ones: for example, the medicine “Amlodipine” => “lodipine” or “anglodipine”, the symptom “diabetic” => “llabetico” and “ colesterol” => “coletrol”. These limitations can be corrected [16] by translating the result obtained into medical vocabulary, since non-existent words can be rescued by an automatic spelling system.

The Google API system, on the other hand, recognized medical words as existing names, such as “Parkinson's” => “parking”, or “obstructions” => “demosthenes”. The system has a high degree of recognition for specific domains of clear speech where a close-talk microphone is used, with noise cancellation and the domain is dictation. The results presented here are similar to those reported by [2]. Google has a medical conversation system available in US English, but not in Spanish.

Systems can be improved through ways: with a significant number of transcribed doctor-patient conversations having been collected, fine-tuning of the acoustic models and language models is possible [1]. Both Kaldi and NeMo are adequate to perform this task.

The precision in entity recognition is high in all systems because common words are hardly transcribed in a medical vocabulary word. The Recall and F1 give a more representative sample of true system performance.

Otherwise, to improve performance could be the recording on different channels to the doctor and the patient, in order to avoid voices overlapping and also the use of appropriate microphones to improve audio quality. The adoption of data augmentation methods for ASR and NLP [17] is another path to follow, the importance lies not only in the deep learning methods chosen but also in the data that trains the systems.

5. Conclusions

The strategy adopted in the language model building, adapted to the domain, a supervised phonetic dictionary and an acoustic model trained with spontaneous speech has allowed the creation of a competitive system for the transcription of medical conversations which

[5] https://github.com/usnistgov/SCTK.git

can be improved in successive developments. To achieve an acceptable degree of recognition of medical entities in terms of F1, it is estimated that a level of Accuracy in speech recognition of 85% could be acceptable. The results of this first evaluation show that the system has potential and that the difficulties can be mitigated in the entity recognition task in the medical domain.

References

[1] Chiu CC, Tripathi A, Chou K, Co C, Jaitly N, Jaunzeikare D, Kannan A, Nguyen P, Sak H, Sankar A, Tansuwan J, Wan N, Wu Y, Zhang X. Speech recognition for medical conversations. arXiv. 2017 Nov, doi: 10.48550/arXiv.1711.07274.

[2] Kodish-Wachs J, Agassi E, Kenny P 3rd, Overhage JM. A systematic comparison of contemporary automatic speech recognition engines for conversational clinical speech. AMIA Annu Symp Proc. 2018 Dec;2018:683-9.

[3] Zuckerman ZE, Starfield B, Hochreiter C, Kovasznay B. Validating the content of pediatric outpatient medical records by means of tape-recording doctor-patient encounters. Pediatrics. 1975 Sep;56(3):407-11.

[4] Romm FJ, Putnam SM. The validity of the medical record. Medical care. 1981 Mar;1:310–5, doi: 10.1097/00005650-198103000-00006.

[5] Renato A, Berinsky H, Daus M, Dachery MF, Jauregui O, Storani F, Gambarte ML, Otero C, Luna D. Design and evaluation of an automatic speech recognition model for clinical notes in spanish in a mobile online environment. Stud Health Technol Inform. 2019 Aug;264:1761-2, doi: 10.3233/SHTI190635.

[6] Xiong W, Droppo J, Huang X, Seide F, Seltzer ML, Stolcke A, Yu D, Zweig G. Toward human parity in conversational speech recognition. IEEE/ACM Trans Audio Speech Lang Process. 2017 Sep;25(12):2410-23, doi: 10.1109/TASLP.2017.2756440.

[7] Lacson RC, Barzilay R, Long WJ. Automatic analysis of medical dialogue in the home hemodialysis domain: structure induction and summarization. J Biomed Inform. 2006 Oct;39(5):541-55, doi: 10.1016/j.jbi.2005.12.009.

[8] Zayats V, Ostendorf M. Giving attention to the unexpected: using prosody innovations in disfluency detection. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers),; 2019 Jun; Minneapolis, Minnesota. 86-95 p, doi: 10.18653/v1/N19-1008.

[9] Vogel AP, Morgan AT. Factors affecting the quality of sound recording for speech and voice analysis. Int J Speech Lang Pathol. 2009 Jan;11(6):431-7, doi: 10.3109/17549500902822189.

[10] Ram A, Prasad R, Khatri C, Venkatesh A, Gabriel R, Liu Q, Nunn J, Hedayatnia B, Cheng M, Nagar A, King E, Bland K, Wartick A, Pan Y, Song H, Jayadevan S, Hwang G, Pettigrue A. Conversational AI: the science behind the alexa prize. arXiv. 2018 Jan:1-18, doi: 10.48550/arXiv.1801.03604.

[11] Finley G, Edwards E, Robinson A, Brenndoerfer M, Sadoughi N, Fone J, Axtmann N, Miller M, Suendermann-Oeft D. An automated medical scribe for documenting clinical encounters. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations; 2018 Jun; New Orleans, Louisiana. 11-15 p, doi: 10.18653/v1/N18-5003.

[12] Han W, Zhang Z, Zhang Y, Yu J, Chiu C-C, Qin J, Gulati A, Pang R, Wu Y. ContextNet: improving convolutional neural networks for automatic speech recognition with global context. arXiv. 2020 May, doi: 10.48550/arXiv.2005.03191.

[13] Quiroz JC, Laranjo L, Kocaballi AB, Berkovsky S, Rezazadegan D, Coiera E. Challenges of developing a digital scribe to reduce clinical documentation burden. NPJ Digit Med. 2019 Nov;2:114, doi: 10.1038/s41746-019-0190-1.

[14] Povey D, Cheng G, Wang Y, Li K, Xu H, Yarmohammadi M, Khudanpur S. Semi-orthogonal low-rank matrix factorization for deep neural networks. Interspeech 2018 Sep: 3743-7, doi: 10.21437/Interspeech.2018-1417.

[15] Ogura M, Haynes M. X-Vector based voice activity detection for multi-genre broadcast speech-to-text. arXiv. 2021 Dec, doi: 10.48550/arXiv.2112.05016.

[16] Mani A, Palaskar S, and Konam S. Towards understanding ASR error correction for medical conversations. Proceedings of the First Workshop on Natural Language Processing for Medical Conversations; 2020 July; 7-11 p, doi: 10.18653/v1/2020.nlpmc-1.2.

[17] Zhang A, Xing L, Zou J, Wu JC. Shifting machine learning for healthcare from development to deployment and from models to data. Nat Biomed Eng. 2022 Dec;6(12):1330-45, doi: 10.1038/s41551-022-00898-y.