Audio De-identification: A New Entity Recognition Task

Ido Cohn, Itay Laish, Genady Beryozkin, Gang Li, Izhak Shafran,
Idan Szpektor, Tzvika Hartman, Avinatan Hassidim, Yossi Matias
Google
Tel Aviv, Israel
{idoc,itaylaish,leebird,tzvika}@google.com

Abstract

Named Entity Recognition (NER) has been mostly studied in the context of written text. Specifically, NER is an important step in de-identification (de-ID) of medical records, many of which are recorded conversations between a patient and a doctor. In such recordings, audio spans with personal information should be redacted, similar to the redaction of sensitive character spans in de-ID for written text. The application of NER in the context of audio de-identification has yet to be fully investigated. To this end, we define the task of audio de-ID, in which audio spans with entity mentions should be detected. We then present our pipeline for this task, which involves Automatic Speech Recognition (ASR), NER on the transcript text, and text-to-audio alignment. Finally, we introduce a novel metric for audio de-ID and a new evaluation benchmark consisting of a large labeled segment of the ***Switchboard*** and ***Fisher*** audio datasets and detail our pipeline's results on it.

1 Introduction

Personal data in general, and clinical records data in particular, is a major driving force in today's scientific research. Despite its abundance, the presence of Personal Health Identifiers (PHI) hinders data availability for researchers. Therefore, data de-identification (de-ID) is a critical component in any plan to make such data available. However, the amount of data involved makes it prohibitively expensive to employ domain experts to tag and redact PHI manually, providing a good opportunity for automatic de-identification tools. Indeed, high performance tools for the de-identification of medical text notes have been developed (Dernoncourt et al., 2017a; Liu et al., 2017).

Due to the rise of tele-medicine (Weinstein et al., 2014), clinical records consist of many other types of data, such as audio conversations (Chiu et al., 2017), scanned documents, video, and images. In this work, we direct our attention towards the task of de-identifying clinical audio data. This task is expected to become increasingly more important, as Machine Learning applications in telemedicine are growing in popularity. Given an input audio stream, the objective is to produce a modified audio stream, where all PHI is redacted, while the rest of the stream is kept unchanged. To the best of our knowledge, de-identifying audio is a new task, requiring a new benchmark.

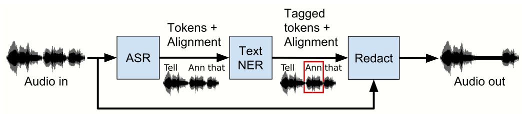

Figure 1: High level audio de-ID pipeline

We define and publish[1] a benchmark consisting of the following: 1. A large labeled subset of the ***Switchboard*** (Godfrey et al., 1992) and ***Fisher*** (David et al., 2004) conversational English audio datasets, denoted as ***SWFI***. 2. A new evaluation metric, measuring how well the PHI words in the input audio were identified and redacted, and how well the rest of the audio was preserved.

To better understand the challenges of the audio de-id task, we evaluate it both end-to-end and by breaking it down and solving it using individual components. Our pipeline (Fig. 1) first produces transcripts from the audio using ASR, proceeds by running text-based NER tagging, and then redacts PHI tokens, using the aligned token boundaries determined by ASR. Our tagger relies on the state-of-the-art techniques for solving the audio NER problem of recognizing entities in audio transcripts (Lample et al., 2016; Ma and Hovy, 2016). We leverage the available Automatic Speech Recognition (ASR) technology, and use its component of alignment back to audio.

[1] https://g.co/audio-ner-annotations-data

Proceedings of NAACL-HLT 2019, pages 197–204
Minneapolis, Minnesota, June 2 - June 7, 2019.

Finally, we evaluate our pipeline and describe its performance, both end-to-end and per-component. Although results on audio are worse than NER performance on text, the pipeline achieves better results than expected despite the compounding pipeline errors. Last, we analyze our performance and provide insights for next steps.

2 Related Work

2.1 NER for Speech

Prior work addressed entity recognition for audio recordings via the *audio NER* task: the detection of entities in the text transcript of the audio input. The majority of these works used a pipeline approach, in which ASR is first applied to the audio and then NER is applied on the noisy textual output of the ASR. These works include discriminative models (Sudoh et al., 2006), incorporating OOV word indicators (Parada et al., 2011), hierarchical structure (Raymond, 2013), and conditional random fields (Hatmi et al., 2013).

Many audio NER works learn from and measure performance on French datasets, such as ***ESTER*** (Galliano et al., 2009) and ***ETAPE*** (Galibert et al., 2014). This may indirectly affect the overall quality of these systems because the ASR component, which is crucial in the pipeline approach but is typically used "off-the-shelf", has lower performance in languages other than English.

An alternative end-to-end approach was proposed by Ghannay et al. (2018), in which the model accepts audio as input and outputs a tagged word sequence which consists of normal words and the NER labels in HTML-like tag encoding. Their model did not attain reasonable performance, perhaps due to the small training set.

We emphasize that both pipeline and end-to-end approaches output tagged word sequences, and do not propagate the recognized entity labels back for redaction on the audio itself, which is the end goal of our proposed audio de-ID task.

2.2 De-identification in the Health Domain

Previous efforts of de-ID in health care focused on redaction of textual medical records. The main approach involves applying NER techniques to the text, including rule-based (Ruch et al., 2000; Neamatullah et al., 2008) and machine learning (Guo et al., 2006; Yang and Garibaldi, 2015) methods.

Adoption of neural network models boosted the performance of NER on text without requiring hand-crafted rules and complex feature engineering (Collobert et al., 2011; Huang et al., 2015; Lample et al., 2016; Ma and Hovy, 2016; Dernoncourt et al., 2017a). Dernoncourt et al. (2017b) applied the model proposed in Lample et al. (2016) to medical de-ID, achieving state-of-the-art performance on the I2B2-2014 (Stubbs and Uzuner, 2015) de-ID challenge dataset. We have chosen this architecture for the NER component of our pipeline method (Section 5).

3 The Audio De-identification Task

The goal of the Audio de-ID task is to convert an input audio stream into a modified audio stream where the PHI words are redacted. In essence, the goal of the task is to limit the ability of a listener to identify the entities of the conversation while leaving as much information as possible in order to keep the audio understandable.

Formally, the input audio stream is a function $A(t)$ of time, that can be transcribed into a sequence of words $W = \{w_j\}$, where w_j is mapped to the time interval in the audio $T_j = [t_j^{start}, t_j^{end})$. We consider each word to be either PHI or non-PHI, and let I denote the set of PHI words $\{j : w_j \text{ is PHI}\}$.

The output of an audio de-ID algorithm is a zero-one redaction function $R(t)$, indicating which parts of the audio stream are to be redacted, where a value of zero indicates PHI information at time t. The redacted audio stream can be obtained by zeroing out the redacted part of the stream, $A_{redacted}(t) = R(t)A(t)$.

To evaluate the performance of a de-ID algorithm, we term w_j as *fully-covered* if $R(t)$ is zero for all $t \in T_j$, and define a corresponding indicator function $covered(w_i)$. This in turn defines the following standard NER metrics for the audio de-ID task:

$$\begin{aligned} TruePositives\ (TP) &= \sum_{j \in I} covered(w_j), \\ FalsePositives\ (FP) &= \sum_{j \notin I} covered(w_j), \\ FalseNegatives\ (FN) &= \sum_{j \in I} 1 - covered(w_j) \end{aligned}$$

$$Precision = \frac{TP}{TP + FP} \quad , \quad Recall = \frac{TP}{TP + FN}$$

Finding the exact time interval corresponding to a word is not a trivial task, while redacting

Dataset	Medium	# Notes	# Tokens	% PHI
I2B2'14 train	Text	521	336,422	3.5
AMC'17 train	Audio	4,629	8,348,899	0.02
SWFI train	Audio	468	710,348	1.8
SWFI test		108	158,923	2.0

Table 1: Dataset statistics for train and test sets, showing the number of notes (written or spoken), token count, and percent of tokens which are PHI.

PHI Labels %	***I2B2'14***	***AMC'17***	***SWFI*** train / test
Name	0.84%	0.12%	0.22% / 0.23%
Age	0.24%	0.01%	0.12% / 0.1%
Date	1.56%	0.03%	0.1% / 0.12%
Hospital	0.28%	0.004%	-
Pharmacy	-	0.01%	-
Organization	0.025%	0.003%	0.48% / 0.59%
Location (General)	0.001%	0.004%	0.24% / 0.29%
State	0.07%	-	0.15% / 0.16%
City	0.08%	0.003%	0.25% / 0.29%
Country	0.02%	-	0.23% / 0.27%
Profession	0.04%	-	0.23% / 0.27%
Holiday	-	-	0.12% / 0.07%
Season	-	-	0.04% / 0.03%

Table 2: Statistics for PHI labels as percent of total tokens per dataset. Tags in **bold** are common to all datasets and are used in Section 7

most of the interval T_j results in a similar de-ID effect as fully covering all the interval. To this end, we extend $covered(w_i)$ into the indicator $\rho\!-\!covered(w_j)$ that is true iff $R(t)$ is zero on at least ρ proportion of interval T_j.

With this indicator function we further extend the aforementioned NER metrics to TP_ρ, FP_ρ, and FN_ρ, and correspondingly define $Recall_\rho$ and $Precision_\rho$. When $\rho = 1$ these metrics equal the strict metrics. When $\rho < 1$ the new metrics determine the quality of the system with respect to redacting at least ρ of each audio interval in PHIs.

We note that the proposed metrics only evaluate the redaction function $R(t)$ on the word intervals.

4 Datasets

To create a benchmark for the audio de-ID task, we use three datasets from two distinct domains: *conversational English* and *medical records*. We summarize the main dataset statistics in Table 1. Importantly, we did not perform text normalization specific to each domain.

Word Type	WER	Well Aligned	Extended Alignment	Shortened Alignment
PHI	41.8	86%	5%	9%
non-PHI	38.3	81%	8%	12%

Table 3: ASR WER and token-audio alignment distribution on sample conversations from the ***SWFI*** dataset.

In the domain of medical datasets, we use ***I2B2'14*** (Stubbs and Uzuner, 2015), which consists of identified textual medical notes with PHI tagging, and the Audio Medical Conversations dataset from (Chiu et al., 2017), denoted ***AMC'17***, which contains de-identified audio of doctor-patient conversations and their corresponding manual transcripts. Processing the ***AMC'17*** conversations was facilitated by the fact that it is a de-identified dataset, which provides us with the locations of the PHI in the audio and the transcripts. Three PHI types: names, dates and ages were redacted, preserving type information, and synthetic data was generated using dictionaries and context-aware rules. First names were drawn from the US Social Security Administration babies names registry[2] and last names were drawn from the Frequently Occurring Surnames list from the 1990's US Census[3]. Human annotators used surrounding context to resolve the other PHI types and filled in fake appropriate identifiers.

Notably, neither of the above-mentioned medical datasets could serve as a benchmark for the audio de-ID task, as ***I2B2'14*** is text-based, and ***AMC'17*** contains only redacted audio conversations and is not publicly available. Therefore, we focused on the conversational English domain, where we generated a combined dataset ***SWFI*** from the ***Switchboard*** (Godfrey et al., 1992) and ***Fisher*** (David et al., 2004) datasets. These datasets include hundreds of conversations in English about a variety of subjects, along with their transcripts. To enable proper training and evaluation for the audio de-ID task, we annotated all 250 ***Switchboard*** conversations, and 326 from ***Fisher***. Annotation included named PHI labels, and the time intervals $T_j = [t_j^{(start)}, t_j^{(end)})$ matching each named PHI back into the audio. This dataset is publicly available[1] to allow for standardized evaluation of novel approaches to this task.

The annotation process began by tokenization of the transcripts provided in both datasets using white-space separators, removing special transcript characters and keeping word capitalization in its original form. Following that, PHI word annotation was performed manually. The results can be seen in Table 2.

[2]https://www.ssa.gov/oact/babynames/
[3]https://www.census.gov/topics/population/genealogy/data/1990_census.htmlcensus_namefiles.html

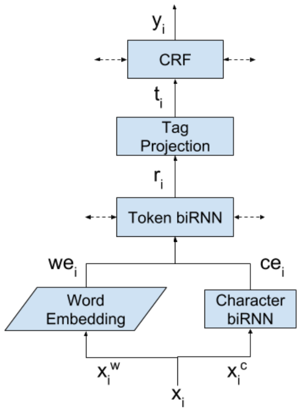

Figure 2: Neural architecture for text de-ID

As performing temporal labeling manually is an arduous process, we opt for a semi-automatic ASR-based procedure. To this end, we determine word start and end times by aligning the manual transcripts to audio intervals. We assess the quality of this semi-automatic labeling scheme using human evaluation. For a random sample of 6 ***SWFI*** conversations (3 ***Switchboard*** and 3 ***Fisher***), we slice the audio according to the aligned interval times per transcript word, and measure both the quality of the transcription, and that of the alignment. Table 3 shows the distribution of alignment errors of the tokens from the sample conversations. These are denoted as good alignment, short (i.e. ASR interval is shorter than actual word) and extended (i.e. interval is longer than expected) where all alignment errors are in the scale of 30-60ms (1-2 audio frames).

5 Pipeline Models

We next describe the models we trained and evaluated to gain insights on the types of challenges this task presents. We chose to use the pipeline approach as an audio de-ID benchmark due to the ubiquity and maturity of the ASR technology, and abundance of training data for text NER. Our pipeline models contain three main components:

1. An ASR system, which transcribes the audio into text.
2. A NER tagger, which tags the transcript with the required labels.
3. An alignment component, which maps each word in the transcript back to its time interval in the audio.

For the ASR component, we use Google Cloud's Speech API[4] with the *command_and_search* model, which gave us the best transcription accuracy on the data. For each conversation, which usually contains two different speakers, we send the entire audio to the service to obtain the transcript. The API also returns alternative hypotheses for the corresponding text and their confidence. We incorporate these alternative hypotheses by taking the top-k ASR hypotheses and feeding them into the next two stages. We then take the logical OR of the detections on all of the hypotheses. Unless stated otherwise, $k = 1$.

For the NER tagger component, our models use the architecture described in Lample et al. (2016), depicted in Fig. 2. This is a neural network model using pre-trained GloVe word embeddings[5] (Pennington et al., 2014) and a character-based bidirectional RNN to generate token embeddings, followed by a bidirectional RNN, tag projection, and CRF layers. We define three models, where each model has a NER tagger trained on a different dataset. The models are:

$\boldsymbol{M_{AMC}}$ – Trained using the training data from the ***AMC'17*** dataset.

$\boldsymbol{M_{SWFI}}$ – Trained using the training data from the ***SWFI*** dataset.

$\boldsymbol{M_{I2B2}}$ – Trained using the training data from the ***I2B2'14*** dataset.

The $\boldsymbol{M_{AMC}}$ and $\boldsymbol{M_{SWFI}}$ models were trained using the conversation transcripts. We use data augmentation in order to increase robustness to ASR errors, in particular to word deletion, insertion, substitution, and inconsistent capitalization. Data augmentation is carried out in several stages. First we create an ASR transcript from the audio, align it back to the reference transcript by minimizing the word-level edit distance, and transfer the labels to the new transcript. For each of the two transcripts, we then generate three additional transcripts by changing word capitalization to camel, lower and upper case. Finally, each of the augmented transcripts is broken down into segments of 20 speaker turns with a step of 10 turns, to resemble the utterance structure of the ASR output. We include the three variants of the $\boldsymbol{M_{SWFI}}$ model: $\boldsymbol{M_{SWFIReg}}$ uses no augmentations,

[4] cloud.google.com/speech-to-text
[5] nlp.stanford.edu/data/glove.6B.zip

$\boldsymbol{M_{SWFI\_MixCase}}$ uses mix-case augmentations only, and $\boldsymbol{M_{SWFI\_MixCase+Asr}}$ uses all mix-case and ASR augmentations.

The $\boldsymbol{M_{I2B2}}$ model is tuned to achieve state-of-the-art results on textual medical notes, such as in Dernoncourt et al. (2017a); Liu et al. (2017). It should be stressed that the model was used as is, without an attempt to adapt it to the domain of ASR output. $\boldsymbol{M_{AMC}}$ and all $\boldsymbol{M_{SWFI}}$ models are both trained on conversational data, and should be better adapted to the task. $\boldsymbol{M_{AMC}}$ is trained on data originating from the medical domain, as opposed to $\boldsymbol{M_{SWFI}}$ models which train on data from the English conversation target domain. This is offset by the fact that $\boldsymbol{M_{AMC}}$ is trained on a significantly larger training set.

Finally, for the alignment component we add a padding hyperparameter allowing a variable number of mismatched frames at either side of the identified intervals. This slack in interval size is used to compensate for alignment errors.

6 Experimental Settings

To test the performance of our models on the audio de-ID task, we conducted a number of experiments, described next. Section 7 then details our results. We report $Recall$, $Precision$, and $F1$ scores for all experiments, which are significantly more informative than accuracy due to a low PHI/non-PHI ratio. We report results on the ***SWFI*** test set using the tags which are shown in bold in Table 2. We evaluate our performance against the coverage threshold $\rho \in [0, 1]$ which is defined in Section 3. Specifically, we focus on type-less metrics, as we care more about the tokens' redaction than their type classification.

Our first experiment evaluates the performance of $\boldsymbol{M_{AMC}}$, $\boldsymbol{M_{SWFI}}$, and $\boldsymbol{M_{I2B2}}$ on the ***SWFI*** test set. First, to decouple their tagging performance from the other pipeline errors, we measure their tagging performance on the manually annotated transcripts (referred to as *NER score*). NER errors may arise due to train-test disparity, where the train and test data are from different domains or different mediums (e.g. text vs. audio), which results in different discriminative models. Additionally, we measure their overall end-to-end score. We analyze the complex behavior of the models' precision by inspecting the coverage distribution of PHI and non-PHI tokens.

Our second experiment evaluates the effect of two significant hyperparameters on pipeline performance using the ***SWFI*** test set:

- The number of alternative hypotheses passed on from the ASR to the NER tagger.
- The amount of padding added around each detection by the alignment component.

Model	NER Recall / Precision / F1	$F1$ (ρ)
$\boldsymbol{M_{I2B2}}$	0.37 / 0.48 / 0.41	0.37 (0.4)
$\boldsymbol{M_{AMC}}$	0.18 / 0.98 / 0.3	0.23 (0.35)
$\boldsymbol{M_{SWFIReg}}$	0.82 / 0.92 / 0.87	0.41 (0.4)
$\boldsymbol{M_{SWFI\_MixCase}}$	0.87 / 0.92 / 0.89	0.46 (0.4)
$\boldsymbol{M_{SWFI\_MixCase+Asr}}$	0.88 / 0.92 / 0.9	0.51 (0.4)

Table 4: NER score of the different NER models, and their end-to-end $F1$ in their optimal choice of ρ.

Error type	Count	% of total
ASR Transcription errors	152	45.24
NER errors	168	50
Alignment errors	14	4.17
Manual Transcription errors	2	0.6

Table 5: Error analysis of a sample of $\boldsymbol{M_{SWFI}}$ FN errors, including errors from all components across the pipeline and even occasional manual transcription errors which contribute to both FP and FN errors.

7 Results

In Table 3 we report the Word Error Rate (WER) of our ASR component on the ***SWFI*** dataset, which was computed by comparing the manual and ASR transcripts of the entire audio. For WER of PHI words, we removed all the non-PHI words from manual ASR transcripts before computing the WER. WER of non-PHI words was computed similarly. We see that both WERs are substantial, and can be thought of as an upper-bound on our pipeline's end-to-end performance.

Next, Table 4 shows the NER and their end-to-end performance of each model for its $F1$ optimal choice of coverage threshold ρ. We can also see that the $\boldsymbol{M_{SWFI}}$ surpasses the others in performance due to its training set being in-domain and in the same medium. Additionally, the $\boldsymbol{M_{SWFI\_MixCase+Asr}}$ variant does not display any advantage over its other variants when running on manual transcripts, but gets significantly better performance on the end-to-end scenario. The difference between NER and end-to-end scores is apparent, and may be attributed to additional pipeline components of ASR and alignment. Interestingly, in the case of $\boldsymbol{M_{SWFIReg}}$, compounding

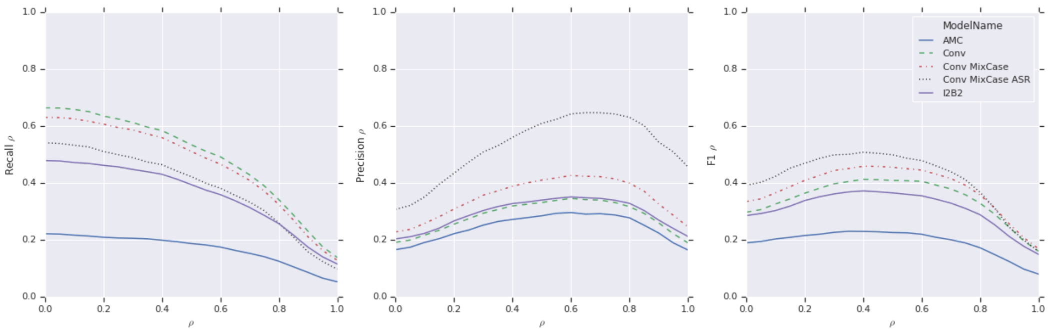

Figure 3: End-to-end performance comparison of NER models – Recall (left), Precision (middle) and F1 (right).

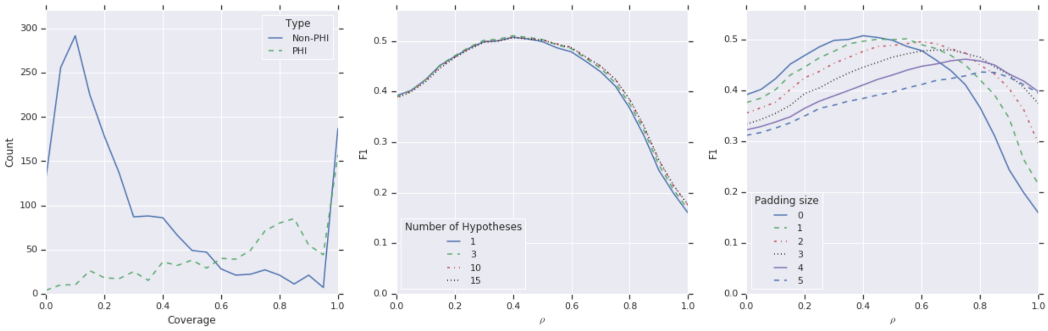

Figure 4: Coverage distribution between PHI and non-PHI tokens (left) and End-to-end performance for different pipeline parameters – number of hypotheses (center) and padding (right).

the WER and alignment error rate from Table 3 and the NER from Table 4 leads to an expected $Recall$ of approximately 0.44, yet the end-to-end $Recall$ at $\rho = 0.5$ is 0.53. This implies a non-trivial co-dependence between errors in the different components of the pipeline.

Figure 3 presents the end-to-end evaluation of the different models with respect to the coverage threshold ρ. As expected, $Recall$ is monotonically non-increasing with respect to the threshold. Meanwhile, $Precision$ (and consequently $F1$) are not monotonic and have more complex behavior. This behavior is due to difference in the distribution of the coverage between PHI and non-PHI, which we see in Figure 4 (left). An interesting insight is that most PHI words have more than half their length redacted by the pipeline while non-PHI words' coverage is bi-modal, one mode close to 0, and the other close to 1. A plausable explanation for this behavior is that the FPs are derived from alignment errors in low coverage, while the high coverage FPs occur due to classification errors, either due to ASR transcription mistakes or due to model NER errors.

Finally, we show the end-to-end evaluation of the pipeline using $\boldsymbol{M_{SWFI\_MixCase+Asr}}$ with different choices of the pipeline parameters. In Figure 4 (center) the performance of the pipeline slightly increases when using additional alternative hypotheses, while a different experiment shows that when using alternative hypotheses with $\boldsymbol{M_{SWFI\_MixCase}}$ performance decreases. This decrease is consistent with the hypotheses' decreasing confidence scores, which can be alleviated with ASR training data but is not addressed by the naive OR approach described in Section 5. This leads us to seek new ways to utilize the additional ASR artifacts, such as the hypotheses confidence scores and speech lattice. In Section 8 we discuss possible directions to improve the pipeline's robustness to ASR errors. Last, Figure 4 (right) shows that the choice of padding size does not improve performance, but rather alters the value of the optimal coverage threshold.

8 Conclusions

We introduced the audio de-ID task, an important prerequisite for protecting privacy when processing sensitive audio datasets in the medical domain as well as other domains. To this end, we created and made available a new test set benchmark derived from annotating the ***Switchboard*** and ***Fisher*** audio datasets. We also presented new metrics for the task, $Recall_\rho$ and $Precision_\rho$, as extensions of standard $Recall$ and $Precision$ where words are considered de-identified when at least a portion ρ of their audio signal is redacted. Finally, we detailed our algorithm for this task, a pipeline approach consisting of three components: ASR, NER on transcripts and a novel alignment from tagged transcripts to audio for the actual redaction.

We showed that ASR performance is the main impedance towards achieving results comparable to text de-ID. In future work, we plan to address this through several directions, including end-to-end de-ID (Ghannay et al., 2018), lattice-based techniques (Ladhak et al., 2016), and diarization and segmentation of the audio as part of the transcription process (Cerva et al., 2013).

9 Acknowledgements

The authors would like to thank Oren Gilon, Shlomo Hoory, Amir Feder, Debby Cohen, Amit Markel, and Ronit Slyper for their generous help in the writing of this paper.

Deidentified clinical records used in this research were provided by the i2b2 National Center for Biomedical Computing funded by U54LM008748 and were originally prepared for the Shared Tasks for Challenges in NLP for Clinical Data organized by Dr. Ozlem Uzuner, i2b2 and SUNY.

References

Petr Cerva, Jan Silovsky, Jindrich Zdansky, Jan Nouza, and Ladislav Seps. 2013. Speaker-adaptive speech recognition using speaker diarization for improved transcription of large spoken archives. *Speech Communication*, 55(10):1033–1046.

Chung-Cheng Chiu, Anshuman Tripathi, Katherine Chou, Chris Co, Navdeep Jaitly, Diana Jaunzeikare, Anjuli Kannan, Patrick Nguyen, Hasim Sak, Ananth Sankar, et al. 2017. Speech recognition for medical conversations. *arXiv preprint arXiv:1711.07274*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *JMLR*, 12(Aug):2493–2537.

Christopher Cieri David, David Miller, and Kevin Walker. 2004. The fisher corpus: a resource for the next generations of speech-to-text. In *in Proceedings 4th International Conference on Language Resources and Evaluation*, pages 69–71.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017a. De-identification of patient notes with recurrent neural networks. *J. Am Med Inform Assoc*, 24(3):596–606.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017b. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Olivier Galibert, Jeremy Leixa, Gilles Adda, Khalid Choukri, and Guillaume Gravier. 2014. The etape speech processing evaluation. In *LREC*, pages 3995–3999. Citeseer.

Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. 2009. The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Tenth Annual Conference of the International Speech Communication Association*.

Sahar Ghannay, Antoine Caubrière, Yannick Estève, Antoine Laurent, and Emmanuel Morin. 2018. End-to-end named entity extraction from speech. *arXiv preprint arXiv:1805.12045*.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, pages 517–520, Washington, DC, USA. IEEE Computer Society.

Yikun Guo, Robert Gaizauskas, Ian Roberts, George Demetriou, Mark Hepple, et al. 2006. Identifying personal health information using support vector machines. In *i2b2 workshop on challenges in natural language processing for clinical data*, pages 10–11. Citeseer.

Mohamed Hatmi, Christine Jacquin, Emmanuel Morin, and Sylvain Meignier. 2013. Named entity recognition in speech transcripts following an extended taxonomy. In *First Workshop on Speech, Language and Audio in Multimedia*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv:1508.01991*.

Faisal Ladhak, Ankur Gandhe, Markus Dreyer, Lambert Mathias, Ariya Rastrow, and Bjrn Hoffmeister. 2016. Latticernn: Recurrent neural networks over lattices. In *Interspeech 2016*, pages 695–699.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *ACL*.

Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *J. Biomed. Inf.*, 75:34–42.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *ACL*.

Ishna Neamatullah, Margaret M Douglass, H Lehman Li-wei, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. 2008. Automated de-identification of free-text medical records. *BMC medical informatics and decision making*, 8(1):32.

Carolina Parada, Mark Dredze, and Frederick Jelinek. 2011. Oov sensitive named-entity recognition in speech. In *Twelfth Annual Conference of the International Speech Communication Association*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Christian Raymond. 2013. Robust tree-structured named entities recognition from speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8475–8479. IEEE.

Patrick Ruch, Robert H Baud, Anne-Marie Rassinoux, Pierrette Bouillon, and Gilbert Robert. 2000. Medical document anonymization with a semantic lexicon. In *Proceedings of the AMIA Symposium*, page 729. American Medical Informatics Association.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *J. Biomed. Inf.*, 58:20–29.

Katsuhito Sudoh, Hajime Tsukada, and Hideki Isozaki. 2006. Incorporating speech recognition confidence into discriminative named entity recognition of speech data. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 617–624. Association for Computational Linguistics.

Ronald S Weinstein, Ana Maria Lopez, Bellal A Joseph, Kristine A Erps, Michael Holcomb, Gail P Barker, and Elizabeth A Krupinski. 2014. Telemedicine, telehealth, and mobile health applications that work: opportunities and barriers. *The American journal of medicine*, 127(3):183–187.

Hui Yang and Jonathan M Garibaldi. 2015. Automatic detection of protected health information from clinic narratives. *Journal of biomedical informatics*, 58:S30–S38.