NUANCE | WHAT'S NEXT BLOG

Healthcare R&D, Innovation & Research

Training automatic speech recognition models with de-identified data

Innovation at Nuance | Employee Guest Blogger

October 3, 2022

De-identification of data used for automatic speech recognition modeling is a critical component in protecting privacy, especially in the medical domain. However, simply removing all sensitive information strongly degrades the recognition accuracy of names, dates, and similar categories. We mitigate this effect by randomly replacing sensitive information, e.g., names by other names, and producing corresponding audio via text-to-speech or snippets from our corpus. Our proposed method almost entirely recovers the general performance loss due to de-identification, and between 50%—90% of the specific accuracy loss for names and dates.

About the author: Martin Flechl is a research scientist at Nuance. His current research interests include speech recognition and text-to-speech. He received his PhD in high-energy physics from Uppsala University, Sweden, in 2010. Prior to joining Nuance in 2019, he worked on machine learning algorithms to analyse data from the LHC particle collider at CERN and led one of the groups which discovered the Higgs boson.

The single most important ingredient for a model based on machine learning is data: The quality and quantity of training material is typically more important for its performance than any other aspect. While

de-identification is a necessary component to protect privacy, it also means artificially decreasing both quantity and quality of our training data. This is particularly true in the medical domain and for products like Dragon Ambient eXperience (DAX) where protecting the privacy of all participants in the medical conversation is a top priority.

For the DAX use case, this means removing essentially all names of patients and providers, a large fraction of dates, and in general a significant amount of numbers, proper names of organizations and places, and much more. Clearly, a model that has never seen any names during training will not be able to accurately recognize names later in the field – while customers rightly expect our products to have that ability. In the paper "End-to-end speech recognition modeling from de-identified data" presented at the Interspeech-2022 conference, we overcome this problem by artificially enriching our training data with words from categories we remove during de-identification, like names.

Method

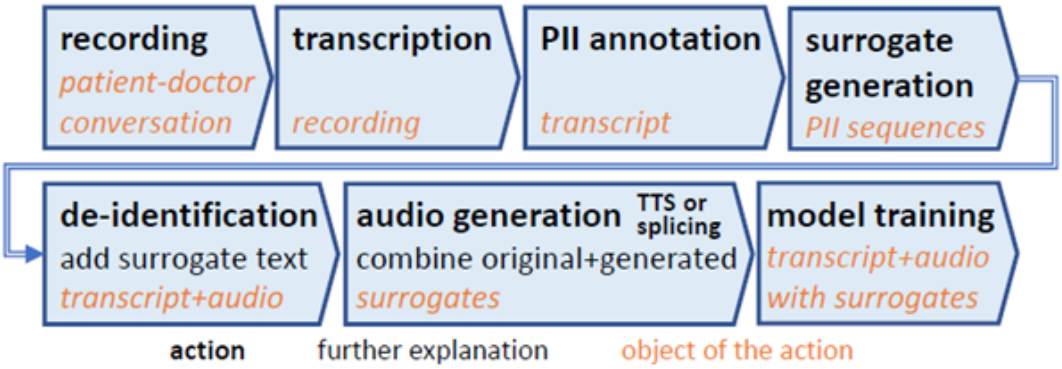

Figure 1: *Method workflow. From the recording of the conversation to the model training with data enhanced by de-identification mitigation strategies.*

Figure 1 shows the workflow of our method. Training data collection starts by recording patient-doctor conversations. The audio recording is then transcribed, and personally identifiable information (PII) is annotated. For these PII sequences, so-called surrogate text is generated: random replacements, e.g., "John" could be replaced by "Michael". Then, during de-identification, PII text is replaced by surrogate text, and the corresponding audio is removed. Finally, audio corresponding to the surrogate text is generated using different methods explained in the following, and the new audio/transcription pairs including surrogates are used for model training instead of the original material.

Each step needs to be carefully executed; for example, if the generated surrogates do not sufficiently cover the range of words expected for recognition in the field, the resulting model will be suboptimal. However, arguably the most demanding step is the audio generation. This is a problem specific to the current state-of-the-art automatic speech recognition models, so-called end-to-end-models: At least in their vanilla form, they exclusively require audio/text pairs for training. Previous architectures implied separately training an acoustic model and a language model. The acoustic model does not suffer from de-identification, and the language model only requires text for training.

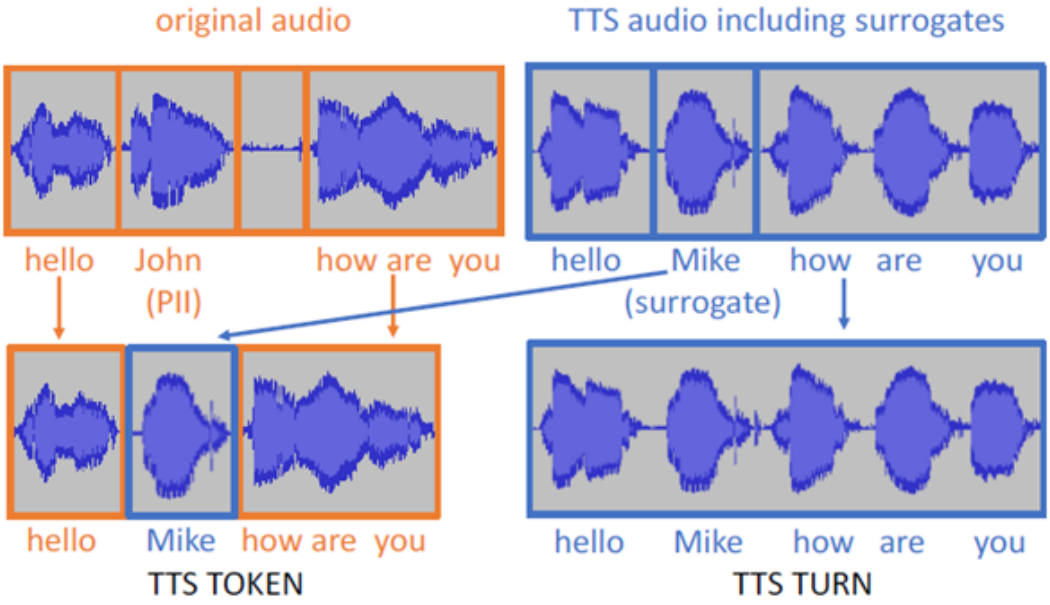

Figure 2: *Audio generation by combining original speech and TTS, or replacing original speech with TTS.*

The first method to generate audio for the surrogate text is based on text-to-speech (TTS) technology. A TTS system is a machine-learning model which has been trained to produce audio output for text input (and metadata, e.g., describing the characteristics of the desired voice). We use TTS as shown in Figure 2: For a given utterance containing PII "hello John how are you", we generate audio for the surrogated version "hello Mike how are you". Then we either simply use the entire TTS utterance (strategy "TTS TURN") or cut out the surrogate ("Mike") and insert it into the surrounding original audio (strategy "TTS TOKEN").

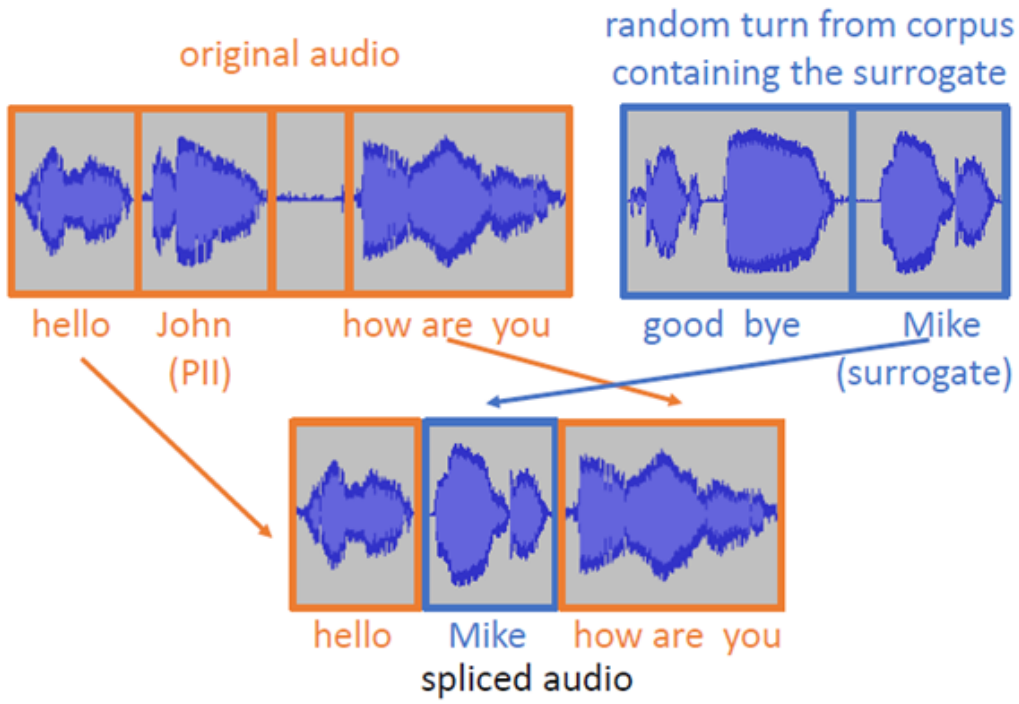

Figure 3: *Audio generation by splicing original speech and snippets from the corpus.*

The second method uses existing audio snippets from our training data corpus. If we require audio for the surrogate text “Mike”, then we search our data for any occurrence of Mike and may find the utterance “good bye Mike”. We then cut out the the audio snippet “Mike” and splice it with surrounding elements from the original audio. Again, two strategies are experimentally tested: either use audio snippets only if they are from the same speaker as in the original audio (“speaker-dependent”) or, if such a snippet cannot be found, fall back to using audio from any speaker (“speaker-preferred”). While the speaker-dependent strategy is more consistent, here it is more likely that we do not find the required word from the same speaker and then cannot use a given utterance for our model training.

Results

We test our methods of enhancing data by training Conformer-Transducer end-to-end models and comparing the results to two baselines: the upper baseline, consisting of identified data; and the lower baseline, where all PII has been removed. The difference between the two baselines is the model performance gap due to de-identification when no mitigation techniques are used. Our methods allow us to recover most of this gap.

In terms of general recognition accuracy, measured by the word error rate, the gap is small and can be entirely recovered using spliced audio and the speaker-preferred strategy, with other methods not far behind. However, our main goal is to improve the recognition of PII-like words. We evaluate F1 scores for different categories to measure the recognition accuracy. For the “date” category, the identified baseline F1 score of 96 drops to 60 after de-identification; all our mitigation methods almost entirely recover the performance loss with the splicing speaker-preferred method scoring highest (F1=94). For names, the performance in terms of F1 score drops dramatically from 77 to 21 due to de-identification. Using the “TTS TURN” strategy, we can recover more than two thirds of that loss (F1=59). Here, the splicing method performs worse because required names often cannot be found in our corpus and hence our training data is severely reduced.

In conclusion, de-identification poses a particular challenge for end-to-end automatic speech recognition. Using our proposed method, the loss in general performance can be recovered. Most importantly, between 50% — 90% of the recognition accuracy drop for categories like names or dates can be recovered as well. Future efforts will focus on combining the splicing and TTS methods, and on improvements to generate speech more closely resembling the surrounding audio in terms of speaker and environmental characteristics.

Tags: Speech recognition, Dragon Ambient eXperience, Research & development