ELEC-E5510 - Literature Study

Hung Nguyen
Joona Sorjonen

Contents

1 What is the topic?

The topic is a speech recognition task using data derived from the Common Voice Corpus, but it has been modified to introduce "gibberish" by making random character changes based on a predefined list. This modified dataset, called "GEO data (Gibberish EsperantO)," consists of increasing degrees of character manipulation. The goal is to develop a robust model that can recognize phonemes even when character alterations are introduced, which can help evaluate model robustness in this recognition task even when facing poisoned data.

2 Why is it a problem?

When phonemes or characters are randomly modified, ASR models are more likely to face ambiguity because they can no longer rely on the predictable speech structure. This is a type of adversarial training - exposing the model to perturbed data during training for stronger robustness. If a model is trained on inputs where phonemes or characters are randomly altered, it can learn to model speech independently. For instance, by training on "gibberish" Esperanto phrases with altered phonemes, the model is "pushed" to develop an understanding of phonetic patterns, rather than rely on sequential dependencies in speech to infer.

Without this robustness, models may fail in practical scenarios with unpredictable noise or phonetic variation, which is a very practical problem. In summary, the goal of this task is to solve this problem by reducing overfitting to "clean" training data and instead encouraging it to learn generalized features of phoneme patterns.

3 What are the methods used in this problem?

To address the challenges in speech recognition, recent literature (Wang, Wang, and Lv, 2019; Prabhavalkar et al., 2024) highlights the following main methodologies:

3.1 Traditional HMM-based Approaches

HMM model speech using a set of states, each corresponds to a probability distribution over the observed features, transition probabilities control the movement between states, capturing temporal dynamics of speech. In terms of architecture, HMM-based models often rely on three different components, the acoustic model which maps speech input to a feature sequence, the pronunciation model which is dictionary mapping various levels of pronunciation, and the language model which maps sequences of characters to coherent transcriptions (Wang, Wang, and Lv, 2019). While this modularity allows for flexibility, it complicates the training process since each component is typically trained independently. In contrast, end-to-end (E2E) (completely neural ASR) models, which are the modern modeling approaches, directly map the input audio to the output text without the need for additional modeling systems (Wang, Wang, and Lv, 2019; Prabhavalkar et al., 2024).

3.2 Explicit Alignment E2E Approaches

This family of approaches model alignments between the encoder output and the target sequence explicitly through a latent variable (Prabhavalkar et al., 2024). They often have predefined rules

or structures (alignment schemas) that dictate how the input and output relate to each other.

These are some of the highlighted approaches:

3.2.1 Connectionist Temporal Classification (CTC)

CTC (Graves et al., 2006) aligns the input sequence with the output sequence by generating multiple valid alignments, then computes the probability of the output by summing over these alignments. A key component in this process is the use of blank tokens, which act as placeholders allowing the model to "pause" on certain frames without generating an output character. This enables CTC to handle varying sequence lengths.

In terms of architecture, CTC utilizes encoder, often a DNN (Prabhavalkar et al., 2024), to process the input and map it into a sequence of encoded representations (Graves et al., 2006). Each frame of this encoded sequence is passed through a softmax layer that outputs probabilities over the possible labels and blank tokens. CTC then marginalizes over all possible alignments to calculate the probability of the correct output sequence given the input (Graves et al., 2006).

3.2.2 Recurrent Neural Network Transducer (RNN-T)

RNN-T (Graves, 2012; Graves, Mohamed, and Hinton, 2013) improves on the CTC by relaxing some of its independence assumptions (Prabhavalkar et al., 2024). Similarly, RNN-T uses a blank token to handle frame transitions without outputting a new label. However, RNN-T can produce multiple labels in a sequence between two blanks, unlike CTC, which only allows a single label per frame, allowing greater flexibility (Graves, 2012). In terms of architecture, RNN-T consists of three parts: an encoder, a prediction network, and a joint network (Graves, 2012). The encoder converts input speech frames into high-level features. The prediction network models previous non-blank output. The joint network combines the encoder and prediction outputs to determine the next symbol.

3.2.3 Recurrent Neural Aligner (RNA)

RNA (Sak et al., 2017) is another generalization of CTT, without an assumption of independence between consecutive tokens. Instead, the previous token is passed as an additional input to the model when calculating the probability of a new token. Like previous approaches, RNA defines a probability distribution over blank-augmented labels, allowing it to output either a blank or non-blank label at each frame.

In RNA, valid alignments consist of sequences with a specific number of blank symbols and match the target labels after removing blanks. The model's posterior probability is computed over these alignments, considering both the previously emitted labels and the frames they correspond to (Graves, 2012). This dual conditioning, unlike in RNN-T, allows RNA to better capture the dependencies between output labels and their timing (Prabhavalkar et al., 2024).

3.3 Implicit Alignment E2E Approaches

A major benefit of the previously looked at explicit methods, is that the encoder can operate using only the previously inputted frames to generate encoded frames. This allows the encoder to be used in a streaming manner i.e process data as it is inputted to the model, without the need to have the full data available beforehand (Prabhavalkar et al., 2024). In applications, where this type

of functionality is not required, Attention-based Encoder-Decoder (AED) (Chorowski et al., 2015) models can be used. Unlike explicit alignment methods, which produce output until the final frame is reached, AED models process the entire input sequence at once (Chorowski et al., 2015). The model does not hold an explicit alignment in the models internal state, instead this alignment is held implicitly with regards to the input sequence, neural network internal state and the model attention weights. A softmax layer is finally used to output probabl for potential outputs.

This approach allows AED models compute the conditional probability of the output sequence without making any assumptions of independence between the input acoustics and model outputs (unlike previous models) (Prabhavalkar et al., 2024).

References

Chorowski, Jan K et al. (2015). "Attention-Based Models for Speech Recognition". In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes et al. Vol. 28. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/1068c6e4c8051cfd4e9ea8072e3189e2-Paper.pdf.

Graves, Alex (2012). *Sequence Transduction with Recurrent Neural Networks*. arXiv: 1211.3711 [cs.NE]. URL: https://arxiv.org/abs/1211.3711.

Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton (2013). "Speech recognition with deep recurrent neural networks". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6645–6649. DOI: 10.1109/ICASSP.2013.6638947.

Graves, Alex et al. (Jan. 2006). "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural 'networks". In: vol. 2006, pp. 369–376. DOI: 10.1145/1143844.1143891.

Prabhavalkar, Rohit et al. (2024). "End-to-End Speech Recognition: A Survey". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32, pp. 325–351. DOI: 10.1109/TASLP.2023.3328283.

Sak, Haşim et al. (2017). "Recurrent Neural Aligner: An Encoder-Decoder Neural Network Model for Sequence to Sequence Mapping". In: *Interspeech 2017*, pp. 1298–1302. DOI: 10.21437/Interspeech.2017-1705.

Wang, Dong, Xiaodong Wang, and Shaohe Lv (2019). "An Overview of End-to-End Automatic Speech Recognition". In: *Symmetry* 11.8. ISSN: 2073-8994. DOI: 10.3390/sym11081018. URL: https://www.mdpi.com/2073-8994/11/8/1018.