Statistical Automatic Speech Recognition

Hernan Andres Gonzalez Gongora
Student number 32223978
hernan-andres.gonzalez-gongora5@etu.univ-lorraine.fr

Abstract

In this project, we tested a statistical speech recognition system with different independent variables. Notably, noise levels and young female data show that our acoustic model struggles to identify digits. Male adult speakers and low noise levels yield the best results regardless of length of sequence. Lastly, we visualized these results to facilitate understanding error rates and how they behave depending on the length of the sequence.

1 Introduction

In our automatic speech recognition project, we were tasked to run four experimental setups with the goal of testing different independent variables and the robustness of our speech recognition system. To do so, we developed a small lexicon and a grammar to recognize sequences of digits. These sequences could be pronounced by different speakers and under different noise levels. In each experimental setup, we computed the word error rate and confidence interval.

The acoustic model used for this work was Pocketsphinx's English acoustic model to which we pass our custom lexicon and our language model uses our grammar to builds its rules. This resulting grammar is also passed to the acoustic model to finalize setting up our custom decoder.

1.1 Lexicon

To create our lexicon, we fetched the provided English lexicon and search for the digit words. This resulted in 13 possible digits with varying degrees of pronunciation

It is key to properly define our lexicon, otherwise certain digits like *oh* would not be properly processed by the ASR system. In one of the first iterations of the system this was the case, which caused performance issues. Additionally, it included the digit *fives* represented as *F AY VZ*. However, this did not change the results whatsoever.

Digit	**Phonetic Representation**
zero	Z IH R OW
zero(2)	Z IY R OW
oh	OW
one	W AH N
one(2)	HH W AH N
two	T UW
three	TH R IY
four	F AO R
five	F AY V
six	S IH K S
seven	S EH V AH N
eight	EY T
nine	N AY N

Table 1: Predefined Lexicon for our ASR system

1.2 JSpeech Grammar

Without properly defining a grammar with its respective rules, the ASR system would be unable to know what rule the decoder must be initialized with. The grammar we created is also contingent upon the lexicon we created, as such both must be consistent.

In our grammar, four main sequences were defined based on the JSpeech grammar format

- **Sequence of length 1**: number
- **Sequence of length 3**: number, number, number
- **Sequence of length 5**: number, number, number, number, number
- **Unknown Sequence**: number+[1]

[1] Similar to RegEx notation, the *plus* sign indicates that one or more digits is valid in this sequence

To properly feed each rule with its corresponding value, an iterable of numbers was defined from one to nine. Two additional values were appended to this iterable, *zero* and *oh*, as they correspond to the number *0*. This also adheres to our lexicon in Table 1

2 Experiments and Results

1. **Experiment 1**
   - Using different language models
2. **Experiment 2**
   - Speaker groups from different ages
3. **Experiment 3**
   - Groups split by gender
4. **Experiment 4**
   - Split by noise level

We can find all of the results in Table 2. Not only was each experiment tested on every sequence, but an overall error rate was computed. This was done by concatenating each one of the reference and hypothesis files and then computing the error rate from this aggregate. An alternative could have been computing the average between error rates, but since each word error rate is computed based on the amount of sequences per file, the output would not be based on the real total amount of sequences.

2.1 Experiment 1

In this set up, we see the highest word error rates are linked with unknown and N-gram sequences. Unlike the other rules, the N-gram rule does not take into account multiple sequences, which may explain that the lowest error rate corresponds to sequences containing one digit. Unfortunately, our confidence interval does not fall within the desired threshold, meaning that we cannot be fully certain of these word error rates.

On the other hand, in plot 1 we observe that the word error rate drops with the unknown sequence variable the more digits there are per sequence. Additionally, the confidence intervals for these results as see in Table 2 are under 5% showcasing that these are results are less uncertain than the ones obtained for the N-gram set up. However, the lowest error rates, and with the least variance, correspond to known sequences. These clearly defined rules yield low error rates with an overall confidence interval under the desired threshold. Even increasing our confidence interval to 99% does not change the confidence in these word error rates.

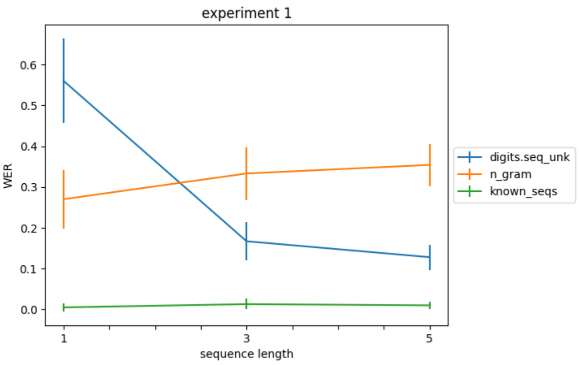

Figure 1: First experimental setup

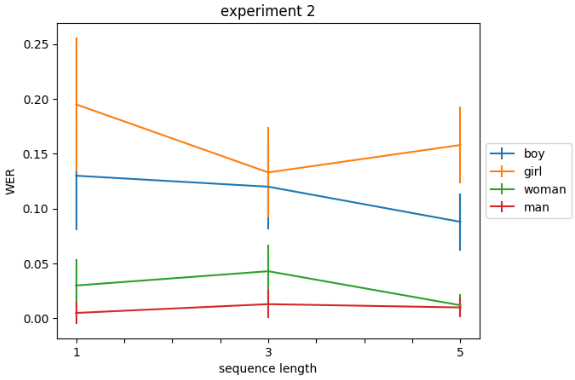

Figure 2: Second experimental setup

Although N-gram rules could be potentially used for sequences with a single digit, known sequences rules are clearly superior and more statistically reliable.

2.2 Experiments 2 and 3

With different speakers that were more statistically relevant, but an interesting phenomenon stood out. The higher error rates in children speakers, especially girl speakers. In fact, the most stable word error rate is the adult male one regardless of sequence as seen in figure 2. This may be stemming from the data the acoustic model was trained on. It is likely that the training data was not sufficiently diverse to account for speakers of other genders and, perhaps, even age ranges. Even so, the results from Table 2 low error rates from adult female speakers.

The higher performance of adult data is visible in figure 3 with adults. It is even likely that the male speakers drove those error rates down and we

	Independent Variable	Seq1	Seq3	Seq5	Overall
1	Unknown Sequence	0.56±0.104	0.167±0.046	0.128±0.031	0.226±0.029
1	N-Gram	0.27±0.072	0.333±0.065	0.354±0.052	0.331±0.036
1	Known Sequence	0.005±0.01	0.013±0.013	0.01±0.009	0.01±0.006
2	Boy	0.13±0.05	0.12±0.039	0.088±0.026	0.106±0.02
2	Girl	0.195±0.061	0.133±0.041	0.158±0.035	0.158±0.025
2	Woman	0.03±0.024	0.043±0.024	0.012±0.01	0.025±0.01
2	Man	0.005±0.01	0.013±0.013	0.01±0.009	0.01±0.006
3	Adults	0.018±0.013	0.028±0.013	0.011±0.007	0.018±0.006
4	SNR05dB	0.59±0.106	0.596±0.088	0.496±0.062	0.546±0.046
4	SNR15dB	0.095±0.043	0.073±0.031	0.052±0.02	0.067±0.016
4	SNR25dB	0.015±0.017	0.023±0.017	0.02±0.012	0.02±0.009
4	SNR35dB	0.005±0.01	0.013±0.013	0.01±0.009	0.01±0.006

Table 2: All experimental results. First column marks which experimental set up the results belong to.

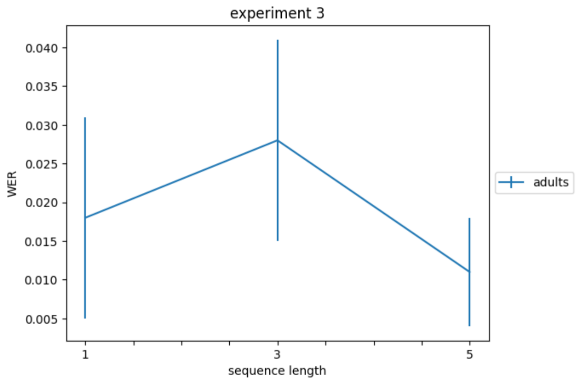

Figure 3: Third experimental setup

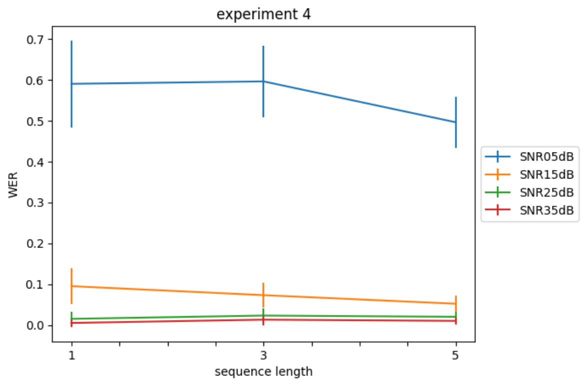

Figure 4: Fourth experimental setup

can see that it is slightly higher than the overall male error rate from experiment 2. Albeit, the confidence interval in both is the same as seen in Table 2. The opposite effect is noticeable as well as the word error rates in overall adult speakers are higher than the isolate male ones. However, as previously stated, the confidence interval remains consistent across the board.

2.3 Experiment 4

Noise levels can have a high impact on ASR systems and ours is not an exception. The highest level of noise yields high error rates with all 3 sequences, bust most notably with sequences 1 and 2. With error rates bordering on 60%. Overall, the model does perform poorly as we can see in figure 4 and the confidence interval for the aggregated results yields a level of certainty in the challenges the model is facing under noisy scenarios.

While the two lowest noise levels, *SNR25dB* and *SNR35dB*, show the lowest error rates with the most reliable confidence intervals, the mid-noise level -i.e. *SNR15dB*- does not fall behind as much. Sequences of length one, do still seem to yield the highest error rate among all sequences under different noise levels.

3 Discussion and Conclusion

Despite some positive results, our experimental set up shows that better defining our unknown rule sequence may be in order. The poor performance is especially notable with shorter sequences. Even if we pass the N-gram model to a sequence of length one, the error rate is still notable at 27%. Further work on preprocessing to clean up noise levels is also advisable, but our results continue to highlight the limitations of acoustic models under noise.

As an extension, we propose changing the unknown sequence rule and comparing the results with state-of-the-art neural models. Even though, they will increase computational time, we presume we will see improvements. Most importantly, however, is to diversify our training data regardless of model. Including more female speakers, especially children, would alleviate the aforementioned challenges.

4 Appendix and Acknowledgements

Training time of all the models takes about 6 minutes. Thanks to Paul Magron for the tutorials and fellow Erasmus students for the support during the development of this project.

Our code can be found in the linked repository[2]

[2] https://github.com/Zappandy/Sphinxbase_speech_recognition