Thesis

# Project Ideas – Research Question (6 relevant phrases)

## What is the minimal number of symbols required to accurately identify a speaker in a group through entropy?

To use Shannon entropy as a marker to rank speakers in a group for identification and analysis purposes. The analysis would determine whether that person’s current speech is significantly departing from that speaker’s norm. This requires a system that is capable of capturing speech, symbolizing it, finding the entropy and returning data quickly. Ultimately the symbolization/alphabet should be minimal such that entropy can be collected with a minimal number of samples needed to return data as quickly as possible. Determining that symbol set is the main focus of research for this project. This symbol set (alphabet) would need to encapsulate the highest valued aspects of speech relative to the desired outcome of analysis for that purpose (i.e. a symbol set would differ for focussing on PWD versus someone with a speech impediment). Understanding specifically what symbols deliver the most amount of information for analysis purposes is crucial.

Potential areas of interest for this system can be people with dementia (PWD) and using this as a way to track whether someone’s speech is suggesting they are showing signs of dementia or keeping track of their speech to elicit warnings to others in the conversation. For PWD these can be symbols of shorter conversational turns, lack of uptake, lack of continuation, or anything that is noticeably out of the norm in regards to speech pausing.

**Minimum symbol set needed for identification of a speaker, and if that speak is differing from their norm**

* 1. Coming up with the novel symbol generation/encapsulation of value/meaning
  2. Minimal set of symbols needed to capture meaning accurately such that identification can be carried out reliably
  3. Analysis of symbol requirements to capture identification from a set
  4. Analysis of complexity from symbol sets
  5. Analysis of symbol requirements needed to calculate entropy that allows identification of a user from a set of users
     1. It then allows for analysing whether that user is departing from their norm state or not
  6. Must be a system that can act FAST, as collection, processing and response should be done ASAP

1. **Analysis of symbol storage requirements and runtime performance as symbol complexity increases**
2. **Implementing more current theory into CalPy**
   1. **Discursis paper Angus mentioned?**
   2. **Using that as part of analysing conversation content to determine symbol requirements?**

# Goals:

**Overarching place in the field:** **Florence Project** - Aiding communication through technology. Analysing, tracking of recurrent patterns in conversation of PWD.

**Overarching project: Andrew’s work** - Detecting change in conversation or detecting when a meaningful pause has been made that shows conversation ending that is context specific (i.e. based on the person speaking), trying to aid in establishing early warning signs of early onset dementia

My project goals: Use symbols to detect changes in speech in relation to pausing through text currently? What about speech? Or possibly a transcript? We want speech to analyse but is that too difficult, should we use a transcript first? Implement Andrews papers.

Related keywords: Dementia, Conversation, Recurrent conversational problems, Analysis and identification of this through text (Potentially transcript), Communication breakdown, symbols?, backchanneling, communication pauses, entropy in language

# Literature Review – What’s been done:

**Estimating Entropy** (Back, Angus, & Wiles, Determining the Number of Samples Required to Estimate Entropy in Natural Sequences, 2018) [[1]](#endnote-1)

**Source:** <https://arxiv.org/pdf/1805.08929.pdf>

**Summary:** Given a set of data to analyse, 1151 symbols are needed to estimate entropy of that data for an alphabet of 3.

**Relevance:** This allows for making correct conclusions about the data when running it through computations, but also provides benchmarks for how many symbols the data should have, avoiding wasting time through unnecessary calculations.

**Fast Entropy** (Back, Angus, & Wiles, Fast Entropy Estimation for Natural Sequences, 2018) <https://arxiv.org/abs/1805.06630>

**Summary:** Method for estimating the Entropy of a given message requiring significantly less samples than the alphabet size (e.g. not relying on slow histogram approaches) with minimal error.

**Relevance:** In trying to determine a speaker from a group, it’s important to be able to determine the entropy for given messages when only small amounts of samples are present, or only requiring small amounts of data to form a close approximation.

**Simple Entropy for Small Datasets** – J. Montalvao, D.G. Silva and R. Attux

**Summary:** Given a small data set, it is not necessary to use a histogram approach to estimate entropy. H can be estimated through estimating the coincidences of characters needing only a small number of samples (relative to the alphabet size).

**Relevance:** This is important for developing a quick estimate of entropy calculation for a given sample set. When trying to quickly determine a speaker from a group and of that speaker is straying significantly from their own norm, entropy calculations should be produced asap

**Conceptual Recurrence Plots** – Daniel Angus: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5887327&isnumber=6180049&tag=1>

**Summary:**

**Relevance:**

**Automated Examination of Dementia Conversation** – Daniel, Janet, Helen: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0144327>

Summary:

Relevance:

**Visualising dementia conversations** – Daniel, Janet, Helen: <https://www.cambridge.org/core/services/aop-cambridge-core/content/view/94641FE0FDBC9F84F38C28B5480F2E13/S0144686X13000640a.pdf/visualising_conversations_between_care_home_staff_and_residents_with_dementia.pdf>

Summary: Using Discursis to analyse the conversational patterns that emerge between PWD's and their carers.

Relevance: This provides documentation of further symbols to look for when estimating entropy and trying to pinpoint the presence of dementia through text or speech, and ways of helping those conversations through specific prompts.

**Human Communication, Quantifying Multi-participant recurrence (MPR) metrics** <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6161608>

Summary:

Relevance:

**Trouble and Repair:** Helen Chenery, Janet Wiles

<https://www.tandfonline.com/doi/pdf/10.1080/026870399402181>

**Summary:** Analysis of common Trouble Indicating Behaviours (TIB's) of patients with Senile Dementia of the Alzheimer's type (SDAT) and the effectiveness of using known repair types to aid in communication with their carers.

**Relevance:** Identifying a structured list of typical PWD breakdowns in communication to aid in potential new paths of symbol creation.

PWD/SDAT (Senile dementia of the Alzheimer's type) conversation traits:

1. Shorter conversational turns,
2. called or regular prompts from the interviewer,
3. increased amount of reference errors,
4. missing elements in conversational turns of SDAT subjects,
5. more topic initiation and unexpected topic shifts (due to failure to continue and repetition of an idea.
6. Incoherent, meaningless or vague
7. Trouble Indicating Behaviours:
   1. Lack of uptake/Lack of continuation
   2. Reprise/Minimal dysfluency
   3. Pauses
   4. Request for repetition
   5. Not requesting specific information
   6. No hypothesis formation

**An analysis of trouble and repair in the natural conversations of people with dementia of the Alzheimer's type**

<https://www-tandfonline-com.ezproxy.library.uq.edu.au/doi/pdf/10.1080/026870399402181?needAccess=true>

**Signals and Systems: MIT Lectures**

<https://www.youtube.com/watch?v=-FHm2pQmiSM>

**Summary:** Looking at complex systems in abstract terms to simplify to just input and output  
**Relevance:** is this just background research or can I reference it as lit review

It abstracts away unnecessary elements of a problem. Looks at signals as input/output and the relationships between them in terms of a function for studying phenomena. Signal is an x-dimensional function (of time or possibly not, just need a function to model our movement with independent and dependent variables), can be multidimensional.

For our research we are looking at something that is both continuous time (CT) signal and a discrete time (DT) signal, CT in the actual verbal input from the speaker, but will be DT in the computational aspect of recording and analysing. How to convert from CT to DT and what do we need to capture and what is the cost for storing everything? What’s the slowest sample rate we can have?

Functions behave as they always will, they’re bound to their output from their input, but signals and systems is looking at manipulating how their information is sent back to us in a new order essentially. So if it’s f(2x) we are getting every second f(x) point, thereby squishing the data down to get half as much. F(-x) will give us a flipped version of what we are looking at, sending the information backwards, reverse order. F(x-250) is making the timeline wait longer to get to the 250 point, you are pushing it back to happen later (further into the x axis). This is essentially looking at how to manipulate functions given only the input and the output to change, how can we modify the data we’re getting back.

Are we actually just trying to see how to approximate a new function from a given function?

**Discursis**: Discursis would be used to potentially map out the pauses in communication

<http://discursis.com/index.php/news-and-research/>

**CalPy** - Software used to aid in this <https://github.com/YvonneYYu/calpy>

# Research Question

3-4 weeks have a research

1. Since entropy is used more as a means of calculating a minimum compression, could there be problems with using it as an index? Could we incorrectly assess something as having less entropy than we think because of bad or few samples?
   1. Think Andrew's work covers this with the min number of samples needed
2. Shannon's paper
   1. So if an entropy of 3.5 is calculated for F2, does that mean 'words' (letters in runs of 3) of length 3 can be encoded into 4 bits min, but for F0 = 4.7, we can only encode the alphabet into 5 bits min? Or that the last letter of that block can be reduced down to 4bits because of redundancy making it less likely some letters will come up?

# Notes

1. Symbols usually letters

# Bibliography

Back, A. D., Angus, D., & Wiles, J. (2018, May 23). Determining the Number of Samples Required to Estimate Entropy in Natural Sequences. 1-6. Brisbane, Queensland, Australia.

Back, A. D., Angus, D., & Wiles, J. (2018, May 17). Fast Entropy Estimation for Natural Sequences. Brisbane, Queensland, Australia.

1. [↑](#endnote-ref-1)