**Thesis**

Evaluation of the information theoretic properties of an acoustic-prosodic speech event class for classifying conversations with application to people with dementia

# Research Questions

1. Selection of a suitable speech event class which will permit information theoretic models to be developed as a means of analytical conversation classification.
2. Implement the Fast Entropy algorithm and apply to conversation analysis using a range of different speech event classes based on prosodic information.
3. Investigate the role of alphabet size and speech class in effectveiness of detecting trouble in conversations. In partcular, What is the minimal number of symbol samples required to estimate entropy accurately for dialogue and monologue
4. What is the minimal symbol set required to produce **accurate** and **distinct** entropy calculations
5. Creating minimal alphabets that can produce complex and extensive entropies allow for accurately identify a speaker in a group through entropy?
6. Can we find high valued/information symbols that help to capture desired results more efficiently?

# Introduction - Motivations (few paragraphs)

### Entropy, metrics, dementia, conversation characterization

Novel insights from the field of entropy have shown methods to quickly and reliably estimate the statistical properties of a given data source with minimal information, allowing for the design of a novel approach into early detection of trouble in communication requiring minimal sample sets for people with neurological diseases like dementia. This paper will explore the use of the Fast Entropy approach to infer change in the meaning in conversational behaviour and how this can be used to measure whether a conversation can be identified as being atypical. In order to understand trouble in communication and provide automated detection, it is necessary to study the key aspects of conversations (metrics , characterization).

Currently approaches for detecting communication breakdowns or Trouble in conversations are done through user training as locating trouble is context specific which makes universal definition/detection hard [citation]. Trouble can then be better characterized as the loss of meaning in a conversation where information exchange can’t be expressed normally [citation]. A loss of meaning presents itself in an atypical fashion through conversational behaviour patterns by affecting the probabilistic structure with which certain speech events naturally occur [citation]. Given that shannon entropy is an established system of measuring the amount information in a given message/language and the recent advancements of the Fast Entropy method, this project will look into automating initial detection techniques of conversational behaviours in communication. If so, can it be done in a computationally efficient way and could meaning then be measured by using the prosodic elements of speech instead of through semantic analysis.

Prosody as a method of extracting meaning outside of the purely semantic form has been shown [Exploiting Acoustic and Syntactic Features for Automatic Prosody Labeling in a Maximum Entropy Framework][2][3]. In the first paper the method used is maximum entropy. Although a system of maximum entropy provides accuracy, it would be an inefficient implementation if put into a real-time entropy estimation system. Given new research from Dr. Back, a new approach can be developed that aims to provide an estimation of entropy quickly while requiring a minimal sample size.

Recent novel insights from statistical analysis has shown the unintended benefits of quick entropy estimation techniques into how meaning can be inferred through certain speech patterns and their use in speech.

To improve the efficiency with which automated speech analysis tools can provide aid in communication difficulties to people suffering from neurological conditions such as dementia, understanding of language and the inherent probabilistic structure behind it is necessary.

communication breakdown to pwd requires bridging the gap between novel insights in statistical analysis to how trouble emerges through language.

To assist in lessening the effects of communication breakdown between a person with dementia

provide fast detection of trouble.

With recent statistical advancements in the fields of entropy and the

In order to develop automated systems to assist in communication difficulties in neurological conditions such as dementa, it is necessary to study the key aspects of conversations (metrics , characterization). Although other speech systems exist, they dont fully meet the requirements of finding trouble in an efficient way without copious data.

In order to understand challenges in creating metrics its necessary to understand t and B to aid with dementia.

The following sections will describe how trouble is defined and found, automated ways of doing so and the most effective methods currently

# Literature Review (6-7-8?)

### Conversational Breakdowns

Conversational breakdowns in Dementia have been extensively researched to establish where exactly trouble starts occurring in conversations between People with Dementia (PWD) and their carers or loved ones. Trouble occurs when meaning can’t be exchanged sufficiently between either speaker [citation] as conversations require both speakers participating [Chenery?]. Since breakdown occurs when exchange of meaning is impaired, it is not possible to know where trouble is happening without also possessing or inferring some expected characteristics of how the conversation should behave. It is this reason why current approaches for detecting trouble are done through user training as locating trouble is a context specific event which makes relying on semantic information alone hard/not possible for accuracy [chenery? citation]. Meaning is built up and into the conversation coming from not only semantic information but other aspects of speech including prosody [Prosody, information structure and evaluation] where the intended meaning can only be inferred as a combination of both the semantic and the prosodic elements as well (e.g. utterance length inferring insistence or impatience as shown by [Prosody, information structure and evaluation]). This means that if a PWD is experiencing a trouble in communication, it will affect the information being conveyed in the conversation. Thus the distribution of how it is used?

Given how prosodic information can change intent it must also possess a probabilistic structure inherent in it, thus if a PWD meaning is impaired, it changes the nature of what and how much information is being produced [citatiooooon]. Detecting trouble requires understanding the inherent meaning in a given context, if the information being delivered does not match the context that shows trouble right?

### Detecting Trouble in Speech

Conversation breakdowns in speech have been well studied to find key/exact points in communication where failure occurs both in normal conversation and conversations involving PWD. Knowing where trouble itself is occuring can be difficult because locating it relies on understanding how the underlying meaning of a message is incongruent with the specific conversation. Trouble Indicating Behaviours (TIB’s) can be a useful marker for locating where trouble occurs in a conversation. TIB’s are defined by Chenery et. al(1995) as conversational tools listeners can use TIB’s to “highlight points of trouble in understanding a message the speaker is intending”. In this case, how a PWD will use them, why it is useful and the types they frequently use. The underlying pattern that is common among TIB’s used by PWD is the change in conversational structure that shows meaning behind these TIB’s to be at ends with what would make sense in the conversation.?

Examples of TIB’s can be requesting for more or specific information or clarity, a reprisal of a message or “verbal behaviours emitted by the speaker indicating difficulties formulating or producing the message, involving sound, syllable and word repetition, pauses and fillers”, aka as minimal disfluency. According to data taken from Chenery et al, PWD are more likely to rely on minimal disfluency or a lack of uptake in the conversation, both being characterized by frequent or disruptive pauses to the conversation.

This make sense as Dementia causes trouble in forming or recollecting memories, making conversation difficult as it generally relies on certain types of information retrieval (i.e. short or long term). Meaning many types of ordinary speech behaviours by a PWD can become TIB’s because of the impacts on speech that neurological degeneration has, thus making following a conversation and requesting specific information difficult.

The reason they are an indicator of trouble!!! Meaning is lost and incongruence between what was said and relevance to the context is made apparent.

What Chenery’s data showed was that all minimal disfluencies shared specific patterns in that they were an indicator of a lack of understanding in the conversation by delivering very little new information?

What this showed was how TIB’s are found, which is an inference into the meaning of a conversation

### Types of Breakdowns in Conversation/Communication

Chenery’s paper shows Senile Dementia of the Alzheimer’s Type patients produce “shorter conversational turns and called for regular prompts from the interviewer”. Refer to “conversational abilities in SDAT’S” for further examples.

### Reliability and Frequency of TIB’s for Detection of Dementia

Chenery’s research shows TIB’s are a reliable marker of , it’s important to ask about the key characteristics of TIB’s. If a TIB is found, how reliable is this an an indicator of someone suffering from a neurological disease like Dementia? Does a TIB carry enough information with it inherently to be useful? or does it require multple uses in the incorrect context. and if they are Also, how often do we need to wait to find them? Are they common?

In this project, we will follow the same approach which is widely used within the computational natural language processing communnity which is to egard language in general, within a probabilistic framework. Hence, wothin the context of a particular conversation, it becomes possible to characterize a conversation has having a particular probabilisic structure. Consequently, breakdowns in covnersation can be interpreted as changes in the probabilistic strucutre. While this is well understood in terms of lexico-sematnic information, we propose to consider this in terms of acustic-prosodic infrmation. Within an information theoretic onctext. Hhence, this requires the consideration of the most apppririate class of symbolic events to use withn the raw speech signal and then to consider issues such as alphabet size, and suitable algorithms to use for rapidly estimatong the information theoretic properties. It is anticpated that the human expert identified TIBs will ultimately concide with the comptuationally identified changes in entropy. as a s

### Frequently Occuring TIB’s

Chenery et al.’s data shows multiple TIB’s, of all the TIB’s relevant to PWD/SDAT’s, minimal disfluency was shown to occur the most frequently. Minimal disfluency is complicated though. It is not only speech pauses but utterance lengths, tone, pitch, intonation, inflection or gaps in speaking. Ultimately all would serve to aid in the final output to produce results that are technically proficient enough to find meaning correctly with very little variance or false readings and be both accurate and distinct.

The main issue that we will examine is the use of computational models to infer changes int he meaning of a conversation in exactly the same way that TIBs are capable of identifying.

Meaning is hard to infer exactly and completely correctly, however, given that pauses are so well documented it makes it a better decision. Meaning is most important here. People who detect them can see when the meaning behind something isn’t right

### Automating Trouble Detection

Given that TIB’s can be found manually and are a reliable, common and relatively frequent indicator of trouble in language[citation], it’s natural to ask if trouble can then be detected through TIB’s automatically by using natural language processing techniques. Previous research in this area includes [An Automated Approach to Examining Conversational Dynamics between People with Dementia and Their Carers], [Computer-based evaluation of AD and MCI patients during a picture description task]

In trying to automate trouble and repair, one technique was using discursis. Doesnt work with speech so doesnt meet criteria. The Florence project has shown potential avenues of aid to help in this area.

### Criteria for usefulness (change this title)

For a system to be able to automatically detect TIB’s in speech, it must take on the role that any given carer would provide for their patients. To ensure correctness and reliability, a necessary criterion is proposed to determine what is valuable and important. This means avoiding false positives and false negatives in both the *detection* of the right TIB, and it’s intended *meaning* (i.e. it is semantically unambiguous enough to rely on).

In this context this system would need to be able to both identify and also. A patient’s history is required to detect any potential deteriorations in speech to establish what is and isn’t a normal mannerism or potential trouble for that patient as speech and speech patterns are culture, context, individual and language specific [citation].

These requirements are not trivial when considering the level of technological rigour these projects must adhere to in terms of correctness and reliability to be useful. It is not enough to meet these criteria sometimes. This means for automation to be of any value, the system must then meet these requirements reliably:

1. Track that patients progress or deterioration relative to previous conversations
2. Reliably detect specific TIB’s that are present (maximal true positives)
3. Reliably ignore TIB’s that are not present (minimal
4. Be context agnostic (Trouble and TIB’s are not culture, context or language specific, but specific to the PWD/SDAT as TIB’s can change with context)
5. Represent accurately what the speaker is actually saying (or indirectly/subconsciously intending/saying, i.e. no meaning present)
6. Act fast for repair techniques to be a plausible implementation

The ~~TIB’s~~ themselves must adhere to a certain set of criteria as well. ~~TIB’s~~ must:

1. Be as semantically unambiguous as possible (if we’ve found the ~~TIB~~, the symbol representing it should be as unambiguous as possible in meaning)
2. Be Common in occurrence
3. Carry enough information to be insightful, meaningful

Characterising a conversation to typical or atypical, however we define it. Classifying conversations

Conv is typical or atypical

Two ways of detecting a trouble speech

Long term goal is monitor own conv, in different situations over time

Speech conversations in dialogue and monlogue will ahve

Establish this: This is in contrast to speaker identification in whcih case they use classifying system to idenifyy speaker on speech pattenrs, this may form part of longer term goals, but it is not part of the main goal currently, which is just to detect whether trouble can be detected

## Entropy

### Prosody and Entropy

Although entropy has been an established method for estimating the average information content from a given source, a less typical use for it is to provide a means of indexing. Although the idea of indexing using entropy isn’t new, new applications are coming around. Indexing the richness of natural language has been proposed to . Given that language is a naturally rich source of information, it

### 

Prosody as a method of extracting meaning outside of the purely semantic form has been shown [Exploiting Acoustic and Syntactic Features for Automatic Prosody Labeling in a Maximum Entropy Framework][2][3]. In the first paper the method used is maximum entropy. Although a system of maximum entropy provides accuracy, it would be an inefficient implementation if put into a real-time entropy estimation system. Given new research from Dr. Back, a new approach can be developed that aims to provide an estimation of entropy quickly while requiring a minimal sample size.

### Classifying Conversational Behaviours

### Computationally Efficient Estimations

### Evaluating Properties

Conversation classification

### Fast Entropy

### 

### Meaning

### Using Shannon Entropy as an Index

Research has shown ways for meaning to be inferred automatically from speech using neural networks, one example identifies a corresponding emotion to what was said, aiding in clarifying exact meaning/intent [neural-paper]. Unfortunately, neural networks often require copious amounts of data to build models effectively, which is something the current research just doesn’t have because of its novelty and because it’s not fully understood yet. Need a method that doesn’t require lots of samples but can also work fast (e.g. something computationally efficient).

One approach used by [Andrew Back, et al.] uses Shannon Entropy as an index for identification of distinct sources of information that is accurate, distinct and fast, requiring minimal samples needed to work [andrew-1, andrew-2, entropy-est]. Using entropy as a means of indexing is not a new approach and has been used in a range of diverse application and contexts from measuring income inequality, to biodiversity, to industrial diversification, to the lexical richness of natural language [pielou-66, shorrocks-80, attaran-87, levitin-93, thoiron-86, ]. Entropy as an index is well researched. However, recent advances in proving methods of fast estimation have shown this to be a viable technique to be implemented in a real-time system [fast entropy est]. can we also provide accuracy? For the purposes of this project, accuracy may not be so critical .. see comments

In order to provide the types of criteria that are required for an automatic, non invasive system like this, it must require a fairly sophisticated system that can differentiate participants as well as track change over time. Here’s the research that shows this as already being done or possible. By assigning values to participants based upon their speech, it’s possible to implement this system in a computationally efficient way. However, we will need to minimize the alphabet size to make sure complexity grows or it demands too much time to be useful. So in order to implement a multi-vector data structure this will need to contain the minimal amount of symbols required to successfully meet the criteria of producing accurate and distinct results.

It is known that a large alphabet size will require a signiifcant amount fo time to obtain a estimate of entropy. Hence, our interest in this porject is to use a small alphabet size such as that found within orosodic information. In partcular, we will be examining pause information .

## Meaning

## Prosody and Meaning Inferrence(?)

Inherent probabilistic structure underlying how people rely on various aspects of prosody in conversation

### Pauses as symbols in speech

Pauses in speech can deliver a few intended or unintended meanings with them. For a speaker to respond in .5 seconds shows little care in that they are not taking the time to think and process the other speaker’s comments, too quick a response shows no thoughtful contemplation over what was said. Too long and it can show the speaker is not properly paying attention and has become disinterested. Because pauses carry meaning it changes the way they are used, and thus could potentially serve as a marker for a conversational tool for a speaker to rely on, thus becoming part of their vocabulary.

It takes a minimum amount of time for a person to listen, take on board and respond accordingly. I think it takes 700ms just to comprehend the sentence that was said?? Given that there is a biological limit, it makes sense why pauses that are too short can come across as rude. But a pause can also occur when a conversation topic has ended and a new topic is being suggested, which will not carry the same time parameters to illicit meaning that the previous pause was able to.

But parameters that make sense for one region won’t make sense for another. Pause meanings can change with each culture and each person, the Japanese have “one of the shortest conversational replies … often answering before the conversational turn is over”[ Econimist, pauses]. This is not meant as a rude gesture, on the contrary as it helps move on the conversation along. Whereas in Finland it is customary to finish sentences with length pauses.

Hence, it is not essential to unerstand why or what is occurring specifically within each type of conersation, howrever, theuse of information theoretic analysis of conversation is expected to enable the classifcation of convesations into “typical” and “atypical”,. This can be in terms of the same agents or in terms of comapring one conversation aginst others within a broader social context. In each case, it becomes possible to then tag conversations which may exhibit TIBs.

## End Goal

### Multi-Vector Analysis

A set of n alphabets that are able to measure changes in n-dimensional aspects of a conversation.

Ultimately this would allow for max identification/precise analysis of speech change

Minimal disfluency is complicated though. It is not only speech pauses but utterance lengths, tone, pitch, intonation, inflection or gaps in speaking, pauses

This is our main objective. It will have as many symbols as possible. This will contain all the most meaningful elements of TIB’s.

### Entropy

Fast Entropy breakthrough

Minimum sample size

### Meaning

Why is meaning important, why haven’t i talked about that

### Symbolizing Speech

In order to find instances of disfluency and ultimately TIB’s, we must look for smaller patterns in speech that we know we can rely on for meaning and correct detection. Certain elements and patterns of speech require symbolization so they can be accurately referred to and measured when they occur. An example being a pause in between speakers, this pause will be a symbol. Symbolization will ultimately allow entropy calculations.

This means symbolizing these TIB’s well enough that they are specific, so not ambiguous for what we are looking for, its common enough, it has an established meaning that is distinct and can be relied for not being hard to distinguish, must carry enough information to be insightful/meaningful, and must represent what the speaker is actually intending.

To make entropy estimation/calculation as efficient as possible, the right symbols need to be chosen. This means rethinking the set of symbols that are being used to analyse speech and try to find new symbols to look for that carry more information than through word analysis alone/that isn’t being implicitly said through a transcript.

These auxiliary forms of communication can be inflection, intonation, deliberate pausing, etc. that all carry inherent meaning with them. Because of this, they are used by speakers in very specific ways that can be thought of as being part of their vocabulary. This serves to identify certain speakers from others.

This can range from looking at the pauses that someone makes in their speech to their pitch, utterance lengths and to tonal shifts. All of these are areas in which we communicate and deliver meaning but are not necessarily well studied for their meaning. Additionally, they can be used to identify certain speakers (as we rely on them as part of our communicational behaviour like we would certain phrases or choice of words, we stick to the ones we like).

Classes and variance each in each - 5-8 pause symbols - how do we measure effectiveness of analysis ~~- need markers of proficiency -~~ can it accurately detect monologue and dialogue

### Information Density of certain types of Trouble/Speech Patterns

To produce a minimal alphabet, it’s necessary to be able to find the patterns in speech which most indicate trouble. To also be able to implement identification then requires unique forms of choice/combination of speech behaviours.

## Pause Code

### Classes of Pause symbols

Previous research into types, occurrence and meaning(?) of pauses in conversation was done by Angus, et al. through the use of the Communications Analytics Lab Python software Calpy and PauseCode. In PauseCode, pauses in speech can either be inner pauses (where a speaker has a significant gap in between speech) or Uptake, the time between a speaker stopping and another speaker picking up the conversation. This form of pause can carry with it a range of potential lengths, varying from 500ms to 5 seconds. This is a form of communication that can change the message being delivered or say something about the speaker’s intention or feelings or engagement in the conversation that isn’t being explicitly said.

We want to symbolize pauses, that means coming up with classes of pauses for the different areas in speech where they occur, and boundaries between which we can quantify them with (e.g. .5 second to 5 seconds). This will help find pauses that are the most likely or meaningful to the conversation and pauses that carry little meaning or information with them. The latter are pauses that can essentially be lumped together as they

So a pause can be after an overtake from speaker B and the return to speaker A, there might be a pause in-between B and A, that pause length can carry various meanings. It could be an indication of someone not paying attention to what was said (very quick return to speaking, i.e. short pause length), it could be a polite pause (.8-9), it could also mean the person is thinking about what was said and taking in the meaning of it to influence what they will say next (longer pause length). This pause class could then carry 3 symbols (possibly).

Google (Response time conversations)

<http://www.speech.kth.se/prod/publications/files/3859.pdf>

<https://www.theatlantic.com/science/archive/2016/01/the-incredible-thing-we-do-during-conversations/422439/>

<http://theconversation.com/awkward-pauses-in-online-calls-make-us-see-people-differently-26073>

<https://www.sciencedirect.com/science/article/pii/S0167639311001580>

<https://www.theatlantic.com/science/archive/2016/01/the-incredible-thing-we-do-during-conversations/422439/>

<https://www.sciencedirect.com/science/article/pii/S1364661315002764>

<https://ac.els-cdn.com/S0095447010000628/1-s2.0-S0095447010000628-main.pdf?_tid=c8148e1b-ddad-4986-9c4f-04fdd454257d&acdnat=1533785742_707951e6b85bedda11decfbb6ea03f68>

<https://www.economist.com/books-and-arts/2017/12/14/the-importance-of-pauses-in-conversation>

How We Talk – NickEnfield

## Calpy

CALPY [github] uses automated signal processing tools to analyse recorded speech and audio processing to detect particular speech patterns. Currently Calpy can produce an automated pitch and pause profile of a given conversation, this allows for extracting data automatically through speech that can be analyzed to find potential TIB’s.

Currently tools like CALPY allow for levels of automatic conversation analysis to extract data in the form of pause and pitch profiles which provides a starting point for initial investigation into how certain characteristics in speech can be measured (and a toolset for symbolizing them). These profiles determine where all pauses (given a upper and lower bound) and pitches in the conversation occur.

Calpy is capable of performing a variety of signal processing tasks suited to speech and audio processing. However it has not yet been extended to cater for symbolic level information theoretic processing, including entropy calculations.

#### What are the areas that Calpy hasn’t covered yet? Gaps between end goal and current state.

To draw conclusive results on trouble detection using an automated process, the speech event classes (define) that are being measured in the conversation need to be as unambiguous as possible. This is hard because TIB’s overlap with normal speech behaviours which means to detect them accurately requires being able to detect when they’re being used correctly or incorrectly, which requires understanding of the underlying/inherent meaning in a given context, the meaning or nature of the conversation and why that meaning is in contrast to the expected usage.

This raises an important question. Can trouble be detected

To make sure TIB’s are detected correctly, general steps are looking at whether misuse has occurred, which means can misuse be detected, if so, can multiple misuses (defined by the speaker themselves and the conversation) be used to accurately identify underlying internal trouble in the conversation and can that trouble be representative of some form of neurological disease like dementia? First steps must make sure that the entropy method being proposed is able to detect and produce enough variance that it could potentially be used as a form of indicator. Given that the method itself is sound, can this method indicate a change in mental health as opposed to a change in author (which is what it originally measured), e.g. one example being Trouble in speech.

Given that trouble oftens shows itself through speech, can that speech be analysed with this method? Entropy and trouble detection make sense together as TIBS are established markers of TROUBLE that emerge through speech/known patterns that occur in speech and entropy uses symbol analysis to infer change (e.g. change in health over a long period of time). This requires that a new alphabet be proposed comprised of elements in speech to track how it is used by a given speaker, which can track change over a period of time (as dementia is a slow process and likely won’t emerge extremely apparent in a single conversation until it’s too late down the track).

they are able to try and establish a variance from the norm can be detected (which is incorrect as usage varies with person so will produce errors), if so, can variance from a user be detected, and if misuse can be established Is it to do with the TIBs as they are already small components or with measuring conversation accurately?

#### 

# Gaps (pages .5)

Detecting trouble, what is left to be done.

No current technique to perform trouble detection quickly or without huge data sets.

Computationally efficient automated trouble detection system in speech

Quick trouble detection, practical theoretical frameworks to build new research on

We have Fast Entropy, Calpy, information into minimum sample sizes, can we build something that is capable of producing quick/efficient entropy estimations that can show if potential trouble is taking place.

Given the recent development of the massive performance improvement from the Fast Entropy method and the Calpy library, can new techniques be developed looking into how trouble could now be detected in a computationally efficient way? Can an alphabet of a given pause class be created that can detect trouble (or at the very least detect atypical) conversations with reliable accuracy. If so, how much redundancy is present in this alphabet and can it be reduced to an optimal minimum set. Size is important in performance as complex or large symbol sets aren’t time efficient and small sets don’t provide enough detail to be useful. Finely grained symbol sets are important to work with to make sure the symbols themselves are not too broad to draw conclusive results from and any results produced can be reliably reproduced.

From there, if an established alphabet that is both correct while minimizing set size has been proven to be possible to meet the criteria of the project, further alphabets can then be investigated to increase the levels of analysis and accuracy performed on a given conversation.

More specifically how much information can be retrieved from conversations to produce reliable estimates that can pick on trouble in establish foundational estimates into samples needed/expected in reality vs what is proposed theoretically. Example, given 10 two-person conversations, what is the level of accuracy that can be produced for identification, and

What is significant in terms of pauses, changes in pauses and use of pauses? That’s what we hope to find out. Can a pause or a set of pauses be varied enough that it can be used as a reliable source of identification and/or speech pattern change (i.e. trouble indicating behaviour)?

Establish what significance looks like in pauses in practical research through testing and data collection.

Given pauses in a conversation, each type of pause can be considered a class, within each class, this gives rise to a potential alphabet of symbols. Class if framed by conv

Many differeny classes of speech events, we must choose a particualr class of speech event, particualr class of pause, and then focus information theory proprties of that class.

Who’s done what about automating trouble and repair

**Research Plan (~2 pages)**

### Given an established criteria to measure success.

#### Overall Goals

The overall goals of this project are to develop an efficient computational methodology for detecting trouble in various natural conversations. The aim of the model is that it will form part of an online, analytic system, capable of producing results within a small time scale, suited eventually to real-time operation. This is in contrast to previous research which has focussed on offline, descriptive systems.

#### Specific Goals:

Goal A0: - Choosing the initial speech behaviour to model/study

Define an initial, varied set of speech event classes to be used as potential candidates to serve as the basis for classifying conversational behaviour. ~~The Suitability and effectiveness of each speech event class will be tested to determine a key class to be used to structure the initial symbolic class.~~

potential key speech pattern components as candidates to use as a way to classify the measure of the information given in a particular conversation and define a key set, after running tests, that will lay the foundations of what the automated system will be looking for in natural conversations.

Specify a range of speech event classes which may be suitable for evaluating as carrying information theoretic properties which will prove effective in classifying conversational behaviour. This will include different pause structure and other prosodic inforromation. Perform a range of statistical tests on available conversations to determine a porposed speech event class to use within a symbolic signal processing architecture.gyd lep

Goal A: - Produce the alphabet from typical conversations and evaluate then refine

Selecting typical conversations from the database to produce an alphabet with the specified speech behaviour. Evaluate how well the symbol class and the alphabet size does in classifying different conversations.

Research potential ‘typical’ or archetypal conversations to use as an initial controlled, synthetic test to determine how well a particular speech class functions as a means of identifying an atypical conversation. Implement the tests through Calpy with a specified alphabet from a chosen speech event class and evaluate the alphabet size and the symbol class with determining classifications of a range of varied conversations. Refine the speech event class and alphabet size.

Selection of archetype conversations to perform analysis on. Specification of a symbolic alphabet within the nominated speech event class. Implementation of the tests using the Calpy Library. Evaluation of the symolb class and alphabet size in classifying different conversations. This may require the use of controlled experimental convrsations.

Refine the speech event clss and alphabet size.

Goal B0:

Investigate the properties oand the behaviour of the algorithm, specifically the Fast Entropy algorithm with various speech evetn classes to determine likely candidates for processing larger sets of conversations. Describe the effectiveness of the ao selected classes and propose tsks to investigate the limits of the approach in classifying conversations.

**Method**

## Goal A0:

#### Establishing Key Speech Event Classes

To classify the conversational behaviour (e.g. speech patterns that might be present that allow it to be classified as being a typical or an atypical conversation) appropriately, a defined set of key speech event classes will need to be specified first to serve as the foundation for the classification. These speech class events will be based on the prosodic elements of language that serve as a basis for delivering and altering underlying linguistic meaning. Examples of speech event classes can include utterance length, pitch, tone, inflections, intonations.

Prosody (Will probably need to define in my lit review why prosody can be used as a good metric) is used because it’s able to deliver meaning while being relatively easy to detect and measure through speech analysis. A range of statistical tests will be carried out on available, recorded conversations to find which speech event class will be initially selected to define the initial set of potential symbols (or possibly alphabets if the chosen class has multiple ways to be classified) to be used as a means of classifying conversational behaviours when analysing conversations automatically using signal processing.

**What will the tests be looking for?** These symbols will be speech event classes and will be chosen based upon their suitability towards the aim of the project and the inherent amount of information carried within them in determining how likely trouble could be occurring in a given conversation.

Suitability in this context could include the relative occurrence (how likely are we to see this event take place), amount of inherent information/information density (what does it mean towards entropy estimation to find this class or a symbol from this class in a conversation, is it meaningful), ease of detection (is it computationally expensive to run available/current algorithms that can reliably detect the given speech event) and whether the tools for detecting it currently exist (if it doesn’t already exist, determine how hard it will be to implement an automated signal processing tool that will accurately detect the given speech event with minimal errors).

#### Establishing a pause classe to measure

To start measuring and applying statistical analysis to conversations, the analysis performed needs to be as precise and solid as possible to make sure whatever data is produced can be relied on in future as an independent event. Given that pauses are a known trouble indicating behaviour, and are quite simple to identify in audio, this produces a good starting candidate for analysis.

[PauseCode, Other Research] defines various types of distinct pauses that exist in speech. These classes can be defined by their occurence between who is speaking before and after the pause occurs. [PauseCodes] defines two distinct classes of pauses as being an Uptake being a pause bracketed by two different speakers, while an Inner Pause is a pause bracketed by the same speaker. Although there could be N\*N many pause classes for N party conversations, only conversations consisting of two parties will be addressed.

Within each pause class will be a distribution of how frequently each pause of a specific length will occur from that class (e.g. a pause of 200ms could occur 25% of the time). ~~Ideally these classes will serve as distinct, different alphabets to be used independently.~~ Each letter/symbol in these alphabets/symbol sets will be determined by a distinct set of pause lengths they are representing (e.g. a letter/symbol could represent 200ms to 250ms), each letter/symbol will occur with a particular frequency. To find these specific pause classes in speech, CALPY will be used to build pause profiles that list the pauses in a given conversation.

### Goal A:

#### Determining an approach to define symbols

Several ways exist to partition data including bayesian approach, max min approach or ranked statistics. While all these processes have their merits it’s important simply at these early stages in this project to gather data in a way that is simple rather than too complex or sophisticated (not establishing correctness first). Essentially the process must be able to establish minimum and maximum bounds for all potential pauses that can be detected and a way to discretize them into symbols that is easy/simple to initially implement.

#### Histograms

Histograms provide a reliable, simple and visual approach to ordering the data and symbolizing it that provides aid in understanding the data for the initial steps in the project. The parameters here will be in finding the right bin size and maximum/minimum bounds. To produce these histograms, CALPY will be used to analyse audio recordings of natural conversations taken by the media.talkbank.org/CABank/ CallFriend/eng- n/

ca.talkbank.org/access/CallHome/ datasets provided by TalkBank project [Carnegie Mellon U and Penn U] and build pause profiles (where the pauses occur in a given recording) to show the general frequency of how often pauses of specific lengths will occur.

Once progress has been made and the information gathered paints more of a picture then further improvements can be made to increase sophistication of symbol creation (e.g. looking at non-equidistant bin sizes can help provide greater detail/sophistication to the symbolization process).

#### Symbol Candidates

The distribution of events will be investigated to find potential, distinct clusters in the data showing how speakers use pauses in conversation and hopefully the best way to cluster these (i.e. ample clustering now to provide better entropy results but also minimum later to improve efficiency/remove redundancy (luck of finding all symbols)).

To make sure clustering is done with as much thought as possible it’s important to know find all the potential meanings for any given class that is being studied (i.e. a long pause can mean reflection or disinterest). This will help later to pick through the data and understand why clusters form themselves around certain areas and if there may be potential markers in the conversation to infer the meaning of this particular symbol ~~(e.g. a single long pause occurring infrequently could show intermittent contemplation, while frequent long pauses would show disinterest or an inability to keep along with the conversation).~~

This would then require a meta analysis of the symbols observing their frequency in relation to each other over certain periods. Secondly the symbol representing it should be accurately identifying what is meant by the speaker.

[include distribution diagram as an example of how its hard]

This will require varying the minimum and maximum length of pauses and the bin sizes used to collect pauses of certain length together to produce several possible ways in which pauses can be symbolized.

To figure out the best parameters will be an iterative process of looking at the raw data and seeing potential ways of clustering. If bin sizes are too large, too many symbols will be produced, conversely if they’re too small there will be too few to be able to measure anything accurately with them.

Also, if the minimum length for a pause is too small then we will be accepting things that aren’t truly pauses in speech but ordinary dips in speech moving from one word to the next. If it’s too long, this will skew the distribution to one side as pauses of that length will likely not occur, and then clustering together many pauses as one symbol if using equidistant bin sizes.

This will require looking through past research to understand types of pauses and their meaning better, which ones are more likely to occur and produce ways to determine how to symbolize data, and iteratively doing this to refine results (maybe 2 or 3 times).

#### Using Fast Entropy and Entropy Estimations

After enough distinct symbol sets have been created, entropy estimations will be done on the set to determine how much variance can be expected from a given symbol set and how changing features in the way it’s clustered changes this. This will be varied depending on how entropy is estimated in the data, for example changing the window size to estimate entropy of n samples, or allowing that window to overlap other windows (to not bias the samples towards the middle of the window size). Depending on the complexity of the analysis will change potential results (specify how).

#### Symbol Set Tests - Measuring Effectiveness

To accurately rank the given but differently produced symbol sets of a single class against each other, multiple standardized criteria tests will be performed on them to measure how well they can identify an atypical conversation given a normal distribution of conversations to build an estimate from.

To make this as controlled as possible synthetic conversations will be produced that can be used as a benchmark for any proposed alphabet. These conversations will have certain pause behaviours present which will need to be addressed by the alphabet as to whether it can indicate a typical conversation from an a-typical one (just a far enough change in variance of use). To understand what it can pick up and what it can’t. It’s important that controlled tests are done first to establish a proof of concept as to what can be delivered or expected from ideal data. This analysis of complexity from symbol sets will determine a good spot between too small to be useful and too complex to be fully utilized.

From there a proto-alphabet can be used to determine potential minimal alphabets and how to change histogram properties and entropy estimations to come up with alphabets that are faster (larger bin size) or more accurate (smaller bin size). Focussing primarily on correctness first then performance/efficiency. The limiting factor in performance being how much time it takes for specific symbols to occur.

Once atypical can be established, the test will look at how atypical detection can vary across multiple distributions and potential atypical variance. Then look at how much accuracy is provided and how much is needed. Then look at given this range of variance, how long it takes to produce each of these estimations, what trade offs may arise between variance and efficiency. This likely will not produce a clear-cut best symbol set but instead produce enough information to be able to inform better decision making and parameter estimation later to guide and refine how symbol sets are produced and what is important.

Further tests will be conducted on actual conversations to see how it performs. Given that this is new research, this will likely need to be done multiple times to establish what success is, how to move towards it, and how it can vary with the variance in data (i.e. what the bounds of success/non-success look like).

#### Evaluate if CALPY needs refinement

After initial evaluations of the effectiveness of the alphabet (and possibly expected results given an alphabet of it’s size (might need to find other research to give an idea what can be expected?), it can be determined if CALPY requires further finely grained potential class identifiers as the alphabets currently don’t deal with symbols that are well defined enough. Investigate how well calpy classifies different pauses initially then evaluate whether calpy requires further advanced algorithms or if the libraries used are good enough to rely on. This will be examined to determine if there is enough rigour/information present in the software to determine pause structures reliably, accurately.

#### Moving towards

Future Goals

# Project Plan - Gantt Chart

2 week blocks - research and deadlines

26 weeks

Include all things that are due, seminar, demo, thesis,

Week 7 should be full draft

# Risk Assessment

Small paragraph on how method could fail, risks of using calpy in new settings where implementation hasnt been tested.

Work will be done on standard laptop, the risk is no additional risk beyond those of standard computer programming or computing.

# References (~10-20, 20-50, don’t do 200)

# Appendices

12 font