**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. Breakdown or change or shift in whats typical

Novel insights from the field of Shannon entropy have shown methods to quickly and reliably estimate the statistical properties of a given data source, allowing for the design of a novel approach into early detection of trouble in communication 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). Include sentence on TIB’s

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 shifts from typical behaviour [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 has be developed that aims to provide an estimation of entropy quickly.

Previous takes

# Literature Review

### Classification (of Conversations) Through Entropy

Online analytical framework. The use of entropy as a means of classifying a source of information is a well known approach, with the Max Entropy classifier being an established approach that is flexible with the amount of language it can handle and providing highly accurate results [Comparative Study of Classification Algorithms used in Sentiment Analysis]. However, a major drawback to Max Entropy estimations are the amount of samples they require to produce estimations and their complexity [Comparative Study of Classification Algorithms used in Sentiment Analysis]. This limits its use in systems that may have large symbol sets symbols thus requiring a large amount of samples needed, one example being using a classifier on conversations. **Why arent the other classifier algorithms possible to use from the paper?** Max Entropy is desired for its flexibility in modelling but requires large sample sizes to implement. Random Forrest is simple, performs well and is accurate but requires recurrent learning with every novel dataset.

A novel approach developed by [Andrew Back, et al.] uses Shannon Entropy as an index for identification of distinct sources of information that is fast, simple and accurate while requiring minimal samples needed to work [andrew-1, andrew-2, entropy-est].

Given this novel classifier, can a real-time classification system be built to utilize the benefits of this system? One potential classification comes through conversational classification where instead of classifying conversations based upon lexico-semantic meaning, classification can instead be done through the prosodic information available in a conversation. PR

This system ultimately could provide a much richer inference into the given meaning of a message because of the amount of information that is presented in acoustic-prosody while also being computationally efficient. To detect a/typical groups in real-time through some behavioural pattern would likely be relevant, however conversations usually require massive datasets to estimate entropy in conversations [citation], which is why prosody is a viable option.

Use of entropy in conversation analysis has been studied previously [Measurement and Classification of Humans and Bots in Internet Chat] which combined two different classifiers so that accuracy and speed could be obtained. Previous research has looked at the use of a entropy system to provide quick estimations given small datasets [Entropy-Driven Dialog for Topic Classification: Detecting and Tackling Uncertainty].

One important application for testing such a system that requires real-time conversational classifiers can be found in cases of neurological degeneration such as dementia. In this scenario online systems could be implemented to listen to a given conversation and determine if their use of speech event classes varied from the norm significantly. Given a small enough data, accurate and fast results could be computed in real-time. This provides a real application for Fast Entropy.

### 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. However, this is a difficult problem as trouble occurs when meaning can’t be exchanged sufficiently between either speaker [Communication Breakdown : A Pragmatics Problem] 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]. 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, an example being prosodic events [citation]. Given the atypical nature with which a loss of meaning brings with it in a conversation, it should be possible to characterize and classify conversations based on specific behavioural conversation patterns.

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 in both their use of lexico-semantic and prosodic choice. 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 - Conversational Classification

Although trouble itself can be hard to find, it has been shown by Chenery et. al(1995) that internal trouble will manifest itself in predictable ways through use of language for PWD, these trouble markers in speech are called Trouble Indicating Behaviours (TIB’s). 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 in the context of the conversation and the types they frequently use can provide markers in a conversation to signal if trouble is potentially occurring and where.

TIB’s themselves come in a variety of representations in language, used by both people with and without dementia. Chenery shows that for PWD, they will most commonly rely on two forms of TIB’s, minimal disfluency and lack of uptake, both being able to be characterized by prosodic patterns. Minimal disfluency can be characterized as “a disorder in the rhythm of speech in which individuals know precisely what they wish to say, but at the time are unable to say it because of involuntary repetition, prolongation, or cessation of sound””verbal behaviours emitted by the speaker indicating difficulties formulating or producing the message, involving sound, syllable and word repetition, pauses and fillers” [2] [3] [4][citation] (e.g. a pause).

Similarly, a lack of uptake is indicated by a speaker not picking up the conversation after the other speaker drops off, leaving an extended pause in the conversation. Both of these events can be detected through the detection of prosody in speech but more importantly both are examples of pauses which are not typical conversational behaviours which can imply trouble. This provides a way for potential classification of speech into what would be typical and atypical conversational behaviours.

Since inferring where meaning is lost is difficult, it can be a much easier problem to simply identify a general marker for trouble instead. TIB’s can be a useful semi-stable prosodic marker for locating where trouble occurs in a conversation as they can generally be represented by the core components of prosody being utterance lengths, tone, pitch, intonation, inflection or gaps in speaking. Given Chenery shows TIB’s as being a common event and a good indicator of underlying dementia given an increase in frequency among PWD makes them information theoretically meaningful. The new challenge then is to know when a TIB is found, is it representing a legitimate breakdown in conversation caused by dementia or a normal occurrence of trouble and repair.

### Automating Trouble Detection

Given that TIB’s can be found manually and are a reliable, common and relatively frequent indicator of trouble in language among PWD [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, using Discursis, looked at the effectiveness of various communication behaviours and its level of engagement. Although this does look at speech it addresses automation of discource analysis, not in detecting trouble.

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

~~Does a TIB carry enough information How often do we need to wait to find them? Are they common?~~

Previous research on establishing an automated entropy-based classification system in dementia has not been covered in the context this project proposes.

Chenery showed that with a neurological disease such as dementia, TIB’s were far more likely to occur to PWD, specifically minimal disfluencies and lack of uptakes. For example once approach is to consider the prosodic information contained within the pause structure of conversations. Given that these types of disfluencies have long pauses attached to them, a reasonable approach is to consider a probabilistic model of the pause structure. In this project its proposed to examine this changes the inherent probabilistic structure of the prosodic nature of the conversation. We expect there is used causing it to deviate from the norm by excess use of pauses. To be able to detect when a conversation goes from typical to atypical, a different probabilistic structure regarding the use of acoustic-prosody is proposed as the intended means by which classification of one conversation type will be defined from another.

Gathering data on possible speech event classes is important then to answer if a TIB can carry enough information with it to serve as a useful symbol.

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

TIB’s is characterized as a loss of meaning

## 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.

Project builty around extension of calpy. Extended to include fast entropy and information theoretic model, entropy, fast entropy, symbolization, pause profile

Using calpy manually extract pause classes and get

Symbolization step related to pause structure will need refinement or work in order to be suitable, automated or manual process

Further work on entropy long term short term entropy classification

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.

### State of the art

Fast Entropy has been established to be a quick and accurate method for Entropy estimation given small sample size relative to other well known classification algorithms that can underpin a real-time automated system of conversational classification using acoustic-prosody. Amongst the most used TIB’s, minimal disfluency and lack of uptake, showed the importance of pauses in conversation as a metric for detecting potential trouble.

TIB’s have been established to be a commonly occurring, meaningful (as shown with research into pauses), and reliable behavioural pattern amongst PWD for detecting potential underlying trouble in a conversation. TIB’s are a good metric for analysis because of the underlying probabilistic structure that can be used to infer trouble in a conversation allowing classification of typical and atypical conversations through the change frequency of multiple TIB’s. ~~This overlaps with what Fast Entropy can implement well given the efficient computation for small symbol sets for each speech event classes and that the classes themselves represent meaningful information.~~ ~~TIB’s also overlap significantly in speech with acoustic-prosodic speech event classes that makes establishing success more likely.~~

Calpy is an established open-source software library designed for building pause and pitch profile through signal processing of audio files. However it has not yet been extended to cater for symbolic level information theoretic processing, including entropy calculations.

# Gaps (pages .5)

Given a recent, novel approach to conversational analysis, can a online system providing real-time entropy estimation work? What don’t we have?

Currently, no research into specifically what speech event classes, how many, symbol size of each class, how symbol size is best estimated, how to cluster, potential accuracy(could maybe fnd this now), will need to be used to implement classification. Pause detection has been done before, as it be used to classify?

Evaluate what potential acoustic-prosodic events are applicable to developing a real-time behavioural classification need to develop a new alphabet

did not have the Fast Entropy approach until now.

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 well conversational classification can be implemented 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.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)?

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.

**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.

Goal C: System evaluation,

**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.

### System Evaluation ~~Criteria for usefulness (change this title)~~

At the end, use this as a way to evaluate the system.

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).

~~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[4] representing it should be as unambiguous as possible in meaning)
2. Be Common in occurrence
3. Carry enough information to be insightful, meaningful

[Exploiting Acoustic and Syntactic Features for Automatic Prosody Labeling in a Maximum Entropy Framework] has provided evidence that shows prosody detection to be a reliable metric as a speech event class.

#### 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