PROJECT PROPOSAL

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

Abstract

Recent advancements in Entropy have allowed for further avenues in how trouble can be detected in communication for online, analytical systems to allow for efficient classification of conversations based on some symbolized speech patterns. This classification will differentiate a given conversation as typical or atypical based on a selected normal set of conversations.

Robert Cochran - 43135132

Robert.Cochran64@gmail.com

PROJECT PROPOSAL

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

INTRODUCTION - MOTIVATIONS

Novel insights from the field of Shannon entropy have produced 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 the entropy of a conversation, this measurement can then be used to identify whether that particular conversation differed from a norm defined by some set of typical conversations. This project uses work by [1] to establish an application for conversational classification and detecting how trouble in communication will manifest itself in predictable ways through the use of language for People with Dementia. These predictable speech patterns are called Trouble Indicating Behaviours because they act as trouble marker in speech. This project proposes to use prosodic elements of speech as a way to analyse the conversations through the lens of that specific behaviour in a conversation (in this case TIB's) vs how that behaviour is normally used to see how its shift from what's typical [2].

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 [3]. Trouble can then be better characterized as the loss of meaning in a conversation where information shifts from typical behaviour. 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. Given that Shannon entropy is an established system of measuring the average amount of information from a given data source 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.

LITERATURE REVIEW

CLASSIFICATION OF CONVERSATIONS THROUGH ENTROPY

An important use of Shannon entropy estimation in conversational analysis is its potential to produce accurate results of a behaviour relative to extract the relative difference of certain behavioural patterns use between a user and the typical.

To produce an online real-time analytical system for implementing a Shannon entropy estimate of some characterizable behaviour can be beneficial but hard as most that provide ample accuracy require large sample sizes, making a real-time system not possible unless estimations can be done more efficiently. A novel approach developed by Back, Angus and Wiles [4] uses Shannon Entropy as an index for classification of some data source that is accurate, simple and fast.

A major drawback to other entropy estimations are the number of samples they require to produce estimations and their complexity [4]. Given the properties of the Fast Entropy method, this project looks into whether a real-time classification system could be built to utilize the benefits of this system. One potential classification comes through conversation where instead of classification based upon lexico-semantic meaning, classification can instead be done through the prosodic information available in a recorded speech [6]. 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 with enough reliability and information in the speech event classes (some specified character of speech), accurate and

fast results could be computed in real-time. This provides a real application for Fast Entropy given the right choice of speech event classes.

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 [3] as conversations require both speakers participating [1].

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 [1] [3]. Trouble can then be better characterized as the loss of meaning in a conversation where information exchange shifts from what is expected. 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. 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 [7] where the intended meaning of the speaker can only be inferred as a combination of both the semantic and the prosodic elements as together (e.g. utterance length inferring insistence or impatience as shown by [7]). 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.

Detecting Trouble in Speech - Conversational Classification

Although trouble itself can be hard to find, it has been shown by [1] 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 as conversational tools listeners can use to "*highlight points of trouble in understanding a message the speaker is intending*" [1, p. 196]. In this case how a PWD will use them in the incorrect context of the conversation and the types they frequently rely on (potentially because of underlying trouble affecting communication) will indicate the underlying trouble.

TIB's themselves come in a variety of representations in language, used by both people with and without dementia. [1] shows that PWD 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 "*verbal behaviours emitted by the speaker indicating difficulties formulating or producing the message, involving sound, syllable and word repetition, pauses and fillers*" [1] [8, p. 1631].

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 are examples of the types of speech event classes being modelled off of prosody but more importantly both are examples of not typical conversational pause behaviours which can imply trouble.

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, 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 [1] shows TIB's as being a common event and a good indicator of underlying dementia (given the significant increase in usage among PWD), this shows it's a good symbol as a reliance on a single prosody can lower entropy and provide a reliable measure of significant shift in entropy from the norm. 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, it's natural to ask if trouble can then be detected through TIB's

automatically by using natural language processing techniques. Previous research on the topic conversational classifiers is broad, although one was found that specifically looks at detecting trouble in call centre conversations using TIB's and prosody [2] which found TIB's to only be reliably detected in controlled settings.

Background research into automation and dementia found only minimal related papers. [9] looked at trying to automate trouble and repair using a discourse analysis tool, Discursis, looked at the effectiveness of various communication behaviours and its level of engagement. Although this does look at speech it addresses automation of discourse analysis, not in detecting trouble [9]. Another paper by [10] looks at automated performance evaluation of Alzheimer's patients taken through a simple cognitive task. Although this paper produced seemingly significant results, the implementation was extremely controlled, only working in very specific conditions that make reimplementation extremely unlikely [10].

Previous research on establishing an automated entropy-based classification system that aims to ensure correctness of results in dementia has not been covered in the context this project proposes.

Reliability and Frequency of TIB's for Detection of Dementia

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.

Calpy

CALPY [11] 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.

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. 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

Currently, no research or system has been put into place specifically for looking at how an entropy classification system using fast entropy has been found. Although [2] was extremely related, their methodology was extremely uncontrolled. The results for [2] showed that prosody was not a reliable indicator of trouble in natural, unscripted conversations. However, when using an actor to express more emotion trouble was detected with much more reliability. [2] explain this by saying natural conversations may have all these emotions but not express them as well. It could have been the results were poor because the behaviour in question was not optimized appropriately to the behaviour in question (in their case anger). There is good evidence for this explanation since [2] describes their aim was not to optimize single classes or focus on one specific feature but rather try to show what a successful approach towards model could be taken. The results from this experiment are not taken too seriously as correctness was not shown to be ensured either through understanding the behaviour correctly to model it or to ensure the models were actually detecting what they should have. They also provided a different setting looked at a

different conversational behaviour of trouble in communication (anger vs dementia). One important aspect looked at in the paper was the use of controlled tests using actors which was a good choice.

For building effective real-time analytical entropy classification systems around Fast Entropy, there must be much more rigour into how the behaviours are chosen, studied and symbolized (including size of symbol set and how clustering was done to form those symbol sets). Pause detection has been done before, as provided good results, 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

RESEARCH PLAN

OVERALL GOALS

The overarching goals of the project are to develop a symbol sets that allow for implementation of a real-time analysis in a computationally efficient way for distinct and varied 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 A1: - Choosing the initial speech behaviour to model

This involves defining an initial, varied set of speech event classes to be used as potential candidates to serve as the basis for classifying conversational behaviour based on prosodic information. 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.

Goal A2: - 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 through Calpy with a varied set of conversations. Evaluate symbol set size, what symbol class is being used.

Goal B:

Evaluate the effectiveness of the Fast Entropy algorithms with various speech event classes and group sets of classes together to increase the effectiveness of analysis. Look at how effective the sets are and whether Calpy requires further improvements or whether conversation classification currently is where it should be.

Goal C: System evaluation,

Evaluate how the system met the criteria at the end and propose further improvements or direction the project could go.

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.

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.

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 CLASS 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.

[12] defines various types of distinct pauses that exist in speech. These classes can be defined by their occurrence between who is speaking before and after the pause occurs. [12]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). 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 (i.e. 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 from open source databases at 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.

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.

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).

Goal B:

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.

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. 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 implemented in [3] 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.

Goal C:

System Evaluation

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

PROJECT PLAN - GANTT CHART

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

Week 6-7: Research into what potential symbol sets will be most effective initially to be a proof of concept (make sure pause is best)

Week 8-9: Implement Histogram tests through Calpy to determine an initial symbol set to start analysis through inner pause

Week 10-11: Run tests measuring how effective Fast Entropy is providing visual results

Week 10-11: Create synthetic conversations to run through the effectiveness of the initial alphabet, refine how it was produced and recreate alphabets continuously

Week 12-13: Start work on implementing Uptake in Calpy

Sem2:

Week 1-2: Implement joint uptake and inner pause symbol set tests

Week 3-4: Evaluate the effectiveness of Calpy and determine if it requires a finer level of analysis.

RISK ASSESSMENT

Work will be done on a standard laptop; the risk is no additional risk beyond those of standard computer programming or computing. Using Calpy in situations where proper testing hasn't been done could mean software failures.

WORKS CITED

[1] C. M. Watson, H. J. Chenery and M. S. Carter, "An analysis of trouble and repair in the natural conversations of people with dementia of the Alzheimer's type," *Aphasiology,* vol. 13, no. 3, pp. 195-218, 1999.

[2] A. Batliner, K. Fischer, R. Huber, J. Spilker and E. Noth, "How To Find Trouble In Communication," *Speech Communication,* vol. 40, pp. 117-143, 2003.

[3] M. H. Verma, "Communication Breakdown : A Pragmatics Problem," Feb. 2013. [Online]. Available: https://pdfs.semanticscholar.org/a11d/237d7af41f494a8b028e0c0846b5df7f22eb.pdf. [Accessed 23 Aug 2018].

[4] A. Back, D. Angus and J. Wiles, "Fast Entropy Estimation for Natural Sequences," 17 May 2018. [Online]. Available: https://arxiv.org/pdf/1805.06630.pdf. [Accessed 1 Aug 2018].

[5] A. Gupte, S. Joshi, P. Gadgul and A. Kadam, "Comparative Study of Classification Algorithms used in Sentiment Analysis," 2014. [Online]. Available: https://pdfs.semanticscholar.org/4667/88e0ba1f608981ca5422ddfb5bfedeef75d0.pdf. [Accessed 23 8 2018].

[6] V. K. R. Sridhar, S. Bangalore and S. S. Narayanan, "Exploiting Acoustic and Syntactic Features for Automatic Prosody Labeling in a Maximum Entropy Framework," May 2008. [Online]. Available: https://ieeexplore.ieee.org/document/4453862/. [Accessed 21 Aug 2018].