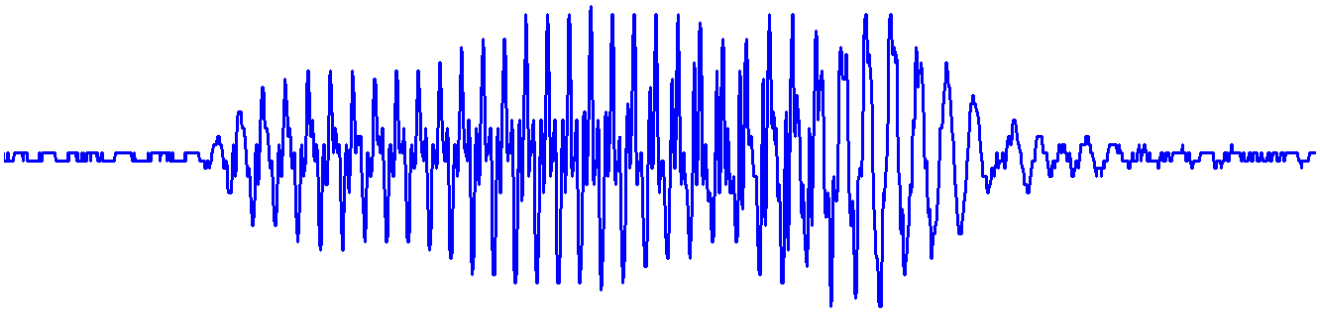


PARALINGUISTIC SPEECH ANALYSER

ECE4095 Final Report

Ruben Michael Bloom
21507252

Significant Contributions

My supervisors, Professor Tom Drummond and Dr. Wai Ho Li, provided invaluable guidance and planning for this project.

MONASH University

Department of Electrical and Computer Systems Engineering

ECE4095 Final Year Project, Semester 1, 2015

Ruben Bloom

Paralinguistic Speech Analysis

Supervisor: Professor Tom Drummond (formerly Dr. Wai Ho Li)

1 Project Aim

1. To identify paralinguistic speech features which contribute to a speaker's *speaking style*, i.e. *how they sound*.
2. To automatically extract these features from recorded speech.
3. To use the developed tools to analyse a test set of speech recordings. This should verify the practical usefulness of the tools and provide insight into the differences between speakers and groups of speakers.

The following features were extracted and analysed: pauses, utterances, pitch statistics, and finality patterns. See below for definitions.

2 What are paralinguistics?

While *linguistics* are what you say, *paralinguistics* are *how you say it*. Paralinguistic speech features are all those aspects that go beyond the words themselves, such as pitch, variations in pitch, pauses, length of utterances, speech rate, umms and ahhs, and more.

Our paralinguistics greatly shape how others perceive us in presentations, interviews, and everyday conversations. They're what separate the boring speakers from the dynamic, exciting, and persuasive! They matter, and this project is about using technology to study them.

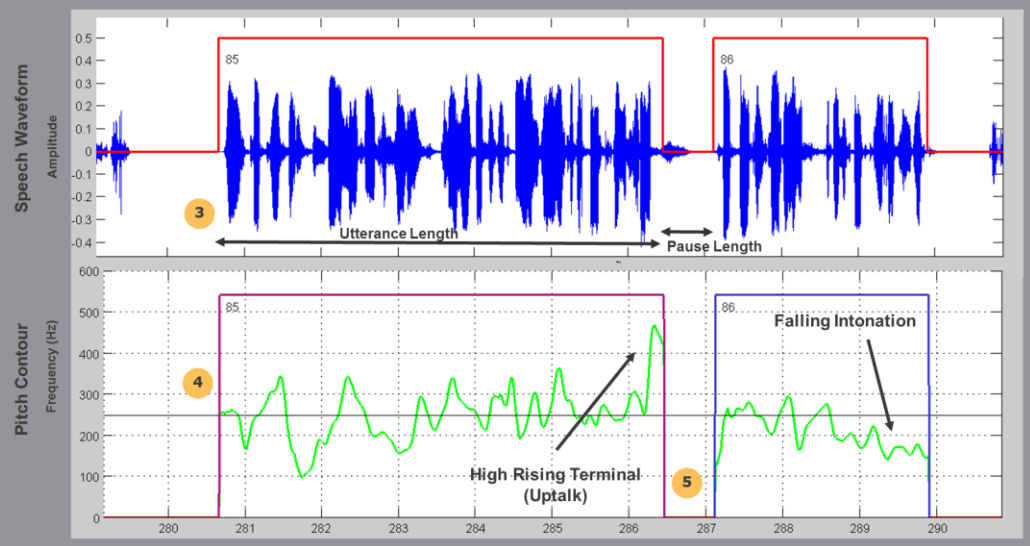


Image: Speech Waveform and corresponding Pitch Contour for two consecutive utterances from one speaker.

3

A Voice Activity Detection (VAD) algorithm identifies the presence and absence of speech in a recording. The output is used to segment the recording into separate utterances and to identify pauses. Additionally, pause and utterance duration statistics are one element of speaking style.

4

The RAPT pitch tracking algorithm is used to calculate the pitch contour from speech recordings. Global pitch statistics such as mean and variance are included in the speaking style profile.

5

The segmentation of utterances is combined with pitch tracking to detect Finality Patterns: the movement of pitch at the end of utterances.

In spoken English, declarative sentences typically end with a drop in pitch (falling intonation), whereas questions end with a rise. However, Australians are known for increasing the pitch even at the end of statement, making them sound like questions. This is known as *High Rising Terminal* (HRT) or "uptalk". Without judging whether uptalk is good or bad, the system detects it.

Example Speaking Style Profiles*

Name	Emily	Jess
Mean f0	249 Hz	167Hz
f0 SD	67 Hz	44Hz
Speech/Pause %	79/21	78/22
Mean Pause Duration	660ms	775ms
Uptalk %	11%	42%
Falling Intonation %	34%	17%

*Real speakers with names changed.

6 SPEECH ANALYSIS RESULTS

4-5 minute speech recordings were analysed from two groups:

- TED talk presentations, where each video had received 1M or more views. They are presumed to be highly charismatic (n =7).
- Oral Presentations from 2nd Undergraduate Psychology Students (n = 7).

TED presenters had higher pitch and pitch variation on average, their pauses were longer, and they paused more of the time.

Student presenters ended 16% of their utterances with High Rising Terminal on average, vs 6% for TED presenters. Students ended 10% of their sentences with Falling Intonation, vs. 26% for TED presenters.

These automatically extracted results match those expected for the difference between experienced, professional speakers compared with the inexperienced: more dynamicness through variation of pitch, well timed pauses, and strong emphatic speech.

Summary

Communication is shaped not just by what it is said, but also how it is said. A new field, computation paralinguistics, has arisen to study all the aspects of speech which go beyond the content. These aspects include pitch, pauses, length of utterances, speech rate, variation of pitch, and others. In combination, these features determine how a speaker sounds, whether they engaging, persuasive, charismatic and a host of other traits. It is the purpose of this project to develop tools for the automatic analysis of *speaking style.*

A review of material in computational paralinguistics and the related field of affective computing was conducted to identify speech features which contribute significantly to how a speaker sounds. Following this, methods for the automatic extraction of these speech features were developed. In the final stage, the tools were tested on three groups of speakers of differing levels of professionalism presenting in formal contexts. The groups were 7 TED Talks with over 1M views each, 7 undergraduate psychology students giving oral presentations, and 6 speakers giving ceremonial wedding speeches.

The analysis with the tools developed revealed expected differences between the groups, validating the practical usefulness of the tools. As expected by previous research into charisma, TED speakers who are presumed to be more charismatic than undergraduate students were found to have higher mean pitch and pitch variation, longer and more pauses, and a higher percentage of sentences ending with emphatic pitch drops. Conversely, the psychology undergraduate students had a high tendency to end their sentences with an upward inflexion, or high rising terminal. Additionally, the analysis revealed dramatic differences between individual speakers, with the bottom five and top five having scores in non-overlapping ranges for all measures. This confirms that the tools are capturing meaningful differences between speakers.

The developed tools for automatic analysis of speaking style will find valuable applications in science and engineering. Tools for quantifying the differences between speakers will allow for greater research into speaker differences and which speaking styles convey which traits. The tools can also be developed into end user products which grant people awareness of their own speaking style and aid them in modifying it in a desired direction. This could be useful for those wishing to improve their public speaking, job interview performance, romantic dates, or even everyday conversation.

Table of Contents

Introduction

"There is no index of character so sure as the voice" - Benjamin Disraeli, British Prime Minister

Many view machines as cold and unfeeling, incapable of understanding human experience, emotional life, or communication. The nascent field of Affective Computing seeks to show that this is not so through the construction of systems which detect human emotion, express emotion, implement emotion, or computationally model emotion. Applications include e-learning, psychological assessment aid, aid in understanding and assisting those with autism or social impairment, robotics, human-computer interaction, emotion research, and market research.

One area to which computing can be effectively applied is the analysis of speaking style. While automatic speech recognition has become widespread, to date it only detects what is said, and not *how it is said*. Considerable information about the mood, personality traits, and intentions of a speaker is contained in how they say things [1]. How things are said is captured in the paralinguistic speech features such as pitch, speech rate, intensity, and rhythm. Collectively, I refer to these features as a speaker's *speaking style*. The features are acoustic properties which can be measured and reported, and they are already widely used in affect[1] detection from speech [2].

There is good reason to measure the paralinguistic features which shape speech. Despite the availability of textual exchange, the overwhelmingly vast bulk of human communication is still spoken. Consequently, understanding our speech and using it effectively is paramount. Those who speak well are deemed more likable, more persuasive, more trustworthy, and receive more attention, giving them an advantage in all domains which involve social interaction, namely work, friends, romance, and family.

Towards this end, the identification, extraction, and analysis of paralinguistic features can be used for the scientific purpose of understanding how speakers differ in their style: it may be easy to hear that two speakers sound very different, but it is harder to qualitatively and quantitatively point to the differences. With the ability to objectively identify and measure differences in speaking style, it becomes possible to answer questions about differences between groups of speakers, between the speaking styles characteristically used in different contexts, e.g. formal and informal, and questions about which speaking styles are most effective, i.e. charismatic, in different situations.

In addition to the scientific purposes, the extraction and analysis of paralinguistic features can be used for an engineering purpose, to create tools which allow users to understand their own speech and help them train towards a desired style, e.g. practice for a public speech, sales pitch, job interview, date, or general conversation skills. For instance, a user with a paralinguistic tool could record their rehearsal of a presentation and discover they are speaking too quickly and with too few pauses. They can continue practicing until they are satisfied – with this tool they have clear, precise feedback.

There are two challenges related to the use of paralinguistic speech features listed above which should be separated: descriptive and normative. The descriptive challenge is that of identifying which speech features are responsible for speaking style, plus the technical difficulties of extracting those features from recorded speech. The normative challenge is that which arises when we

[1] Affect being the technical term for phenomena including emotions and moods.

consider charisma and want to prescribe which speaking styles are better than others in a given context.

This project aims to address both the descriptive and normative challenges, at least in part. Paralinguistic features believed to constitute speaking style were identified, and software was written to extract them from recorded speech waveform. The extraction of features was run on three sets of speakers, with the aim of showing that the paralinguistic features identified capture both differences between the speakers and between groups of speakers. Since one group of speakers consisted of accomplished professionals giving TED Talks with millions of views, and another of 2nd year Psychology undergraduates giving oral presentations, it is assumed that the speaking style of the former group is characteristic of the charismatic speech towards which a speaker would aim. Thus, by describing the speaking style of an a priori charismatic group, an answer can be made towards the normative question of which speaking styles are ideal for given purposes.

Description of the System and Project

This project has two parts: i) Software written for the automatic extraction of paralinguistic speech features from a speech recording, ii) analysis of speakers falling into three different groups using the system developed.

Figure 1 depicts the operation of the software for extracting the paralinguistic features. One or more speech recordings are read into the system and features are extracted. More basic features are used in combination to extract higher-level features, e.g. pitch contours together with voice activity detection and segmentation allows for the classification of finality pattern. The features extracted are discussed at length in the Speech Features section.

Speech by speakers falling into the following three groups were analysed with the software developed:

- 2nd Year Psychology Undergraduate students giving oral presentations on developmental genetic disorders. (n = 7)
- Presentations given TED.com with over a 1M views each. (n = 7)
- Ceremonial speeches given as part of the wedding ceremony of the author by his friends (n = 6)

Complete details of the speakers, recordings, and reasons for their selection are given the Speech Corpora section.

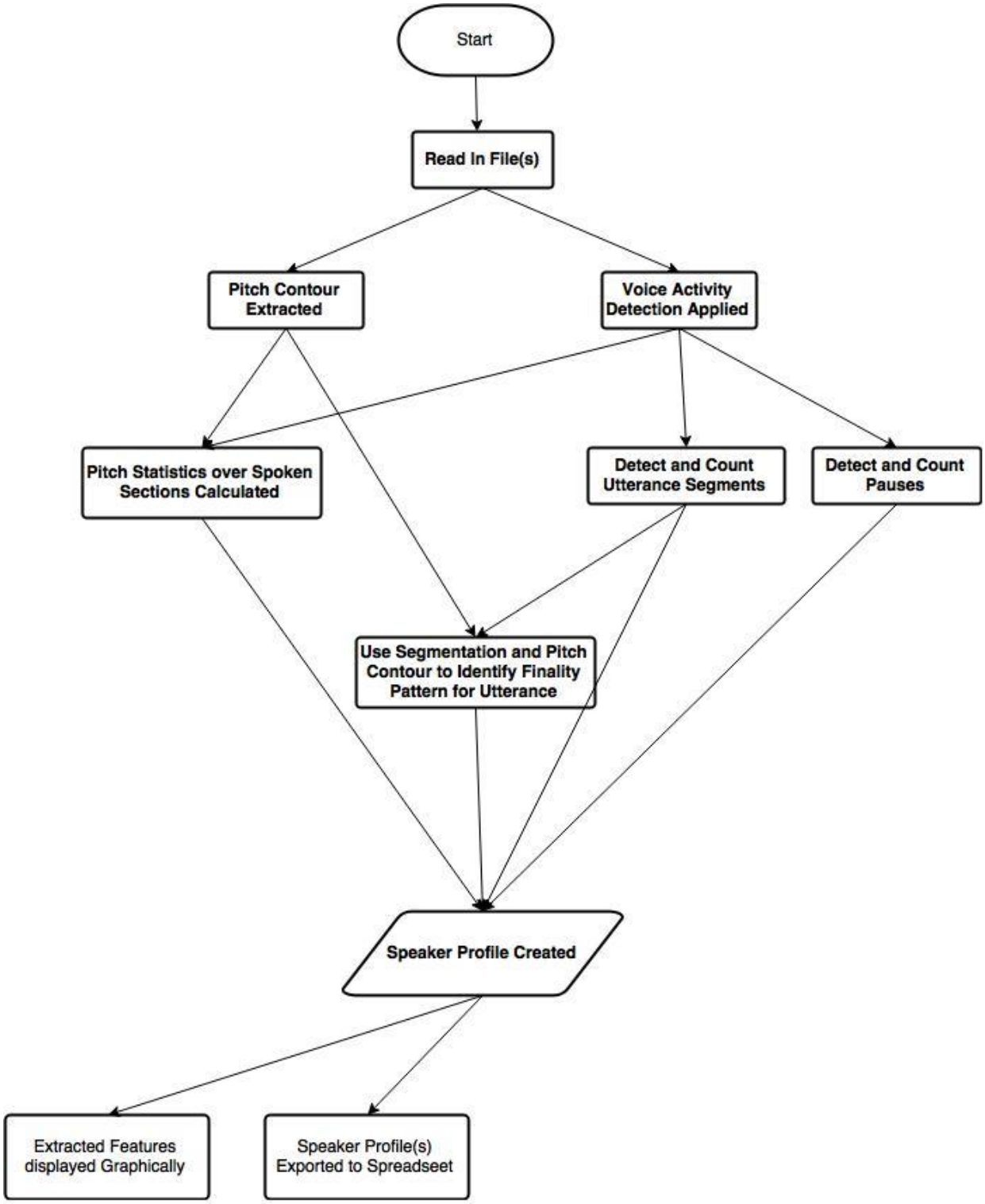


Figure 1 Diagram of System Components

Background

Affective Computing

The origin this project comes from the recently founded field of Affective Computing. An understanding of the field gives an appreciation of the space within this project is created.

Affective Computing research to date has used facial expression, speech, gestures, and physiologically monitoring to detect affective[2] state. Other branches examine detecting conveyed emotion in text. Of relevance to this project is the use of speech features as a source of information about the speaker's state and traits.

The main components of affect detection are the selection and extraction of speech features, and the use of features to train machine learning models which are used for classification. Acquiring suitable data for testing and training is difficult, with the types of emotions and labelling requiring careful selection.

Figure 2 shows paralinguistics speech features used in the detection of various emotions.

	Joy	Boredom	Neutral	Sadness	Anger	Fear	Surprise	Stress	Depression	Happiness	Disgust	Annoyance	Frustration	Anxiety	Dislike
Pitch	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Intensity	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Rhythm	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
Formants		✓	✓	✓	✓	✓	✓	✓	✓	✓				✓	✓
Cross sectional Areas								✓							
MFCC		✓	✓	✓	✓	✓	✓	✓		✓					✓
LFPC				✓	✓	✓	✓			✓					✓
LPC	✓		✓	✓	✓	✓	✓				✓				
Spectral-band Intensity			✓	✓	✓	✓		✓	✓	✓	✓				
Cepstral Coefficients												✓	✓		
Voice Quality Parameters		✓	✓	✓	✓					✓				✓	

Figure 2 Paralinguistic Features used in the detection emotion. Table from Vincialrelli and Mohammadi, 2010 [3]

Computational Paralinguistics

Emerging recently [4], the field of Computational Paralinguistics seeks to extend the work of Affective Computing and use paralinguistic speech information to gather information about a speaker beyond emotion, such as personality [5] and perceived level of charisma. Rightly considered, affective computing using speech is a subfield within computational paralinguistics.

It is an implicit assumption within the research and in this paper that the same speech features which communicate affective state are also the features of speech which are generally salient to people and constitute the perceived speaking style. For example, since a person's pitch informs us

2

about whether they are angry or sad, we can assume that humans are similarly attuned to the variation of pitch for other purposes, making it an important element of speaking style in general.

Speech Features and Charisma Correlates

The goal of this project is to identify and extract paralinguistic features contributing to speaking style. In line with this are efforts by other researchers to find the acoustic correlates of charisma. Charisma encapsulates the vocal traits which make a speaker seem interesting, persuasive, engaging, inspiring, confident, believable, and enjoyable to listen to. Charismatic speakers engage their audience and communicate effectively. Correctly considered, charisma is employment of charismatic speaking styles, which are a subset of all speaking styles.

In these terms, the goal of this project is 1) to identify speaking styles and extract them from speech recordings, and 2) to identify which speaking styles are charismatic in a given context. Accordingly, it is appropriate to review the small body of research on explored the acoustic correlates of charisma.

Rosenberg and Hirschberg [6] had eight subjects rate 45 speech segments from American political speakers on statements about the speaker. Statements were of the form “The speaker is X” where X is one of: charismatic, angry, spontaneous, passionate, desperate, confident, and the like. High inter-rater consistency was found. Statements about enthusiasm, charm, persuasiveness, passion, and convincingness were highly correlated with the statement about charisma.

Rosenberg and Hirschberg analysed the paralinguistic features of the speech samples and found that mean, standard deviation, and maximum f0 all positively influenced the charisma ratings ($p<.001$). Minimum f0 was significant at $p = 0.049$. Speech rate in syllables per second was positively correlated with $p = 0.085$. Faster speaking indicated higher charisma.

Strangert [7] confirmed the results from Rosenberg and Hirschberg. Strangert had subjects rate speech samples from Swedish parliamentary debates on statements concerning whether the speaker was: insecure, hesitant, monotonous, aggressive, accusing, agitating, objective, trustworthy, humble, expressive, powerful, involved, and “the speaker is all in all a good speaker”. Expressiveness, power, and involvedness were found to be positively correlated with charisma.

Strangert found that the highest rated speaker had a higher mean and more varying f0, faster speech rate, and shorter pauses than the lowest rated speaker. These results of dynamic speech matched the correlation between ratings of expressiveness, involved, and agitating with charisma.

Speaker	RH	RL
F0 (Hz)		
mean	138	113
standard deviation	28	12
minimum	75	77
maximum	233	173
range	158	96
Duration (sec)		
total speech sample	33.93	36.62
total pause	4.97	9.05
mean pause*	.41	.57
pause-to-speech ratio	.15	.25
speech rate (syll/sec)	4.27	3.69
articulation rate (syll/sec)**	5.01	4.90

* N = 12 and 16, respectively, for RH and RL
** Calculated after subtraction of pause durations

Figure 3 Comparison of speech features for Rated Highest and Rated Lowest speakers in Strangert 2005

In 2008, Strangert, aided by Gustafson [8], continued her work. In this study, temporal features such as pause duration and speech rate were found to have insignificant and weak correlations. f0 range was found again to be positively correlated with charisma; disfluencies such as repetitions, repairs, and slips of the tongue were negative correlated.

To further test the correlation results, Strangert and Gustafson modified the speech recordings of the least charismatic speaker to increase pitch range, remove disfluencies, and increase speech rate before presenting these modified samples to new the subjects. The subjects rated the modified versions as more charismatic than the unaltered version, verifying the correlation and suggesting that a speaker can improve their charisma level with these modifications.

These studies show that pitch, speech rate, pauses, and disfluencies are all paralinguistic features which strongly influence how a speaker is perceived. These features are excellent candidates for being elements of a speaker's speaking style.

Software Tools

The software for this project was implemented in Matlab.

Initial research for the project included surveying existing tools for speech analysis and selecting which to use. The University of Reading has a page listing available tools [9], among which Speech Filing System [10] and Praat [11] were chosen for consideration. SFS and Praat were chosen for being advanced tools for researchers with good documentation and large user base.

These toolkits implement visualisation and processing of speech waveforms including filtering and spectral analysis. They do not include higher-level algorithms to detect features of interest such as speech rate or filled pause detection.

An important consideration was how these tools could be interfaced with and their functions included in the system. Both include a scripting language.

Advantages of each are:
Speech Filing System
+ Interfaces with Matlab
+ Simpler scripting language

Praat
+ Open Source
+ Very widely used

In the final system, SFS, but not Praat, was used for its implementation of the RAPT pitch tracking algorithm [12]. Other audio processing functions were not needed or sufficiently simple to implement in Matlab. Lastly, although SFS's implementation of the RAPT algorithm was used, later a Matlab implementation was found which could be used in its place.

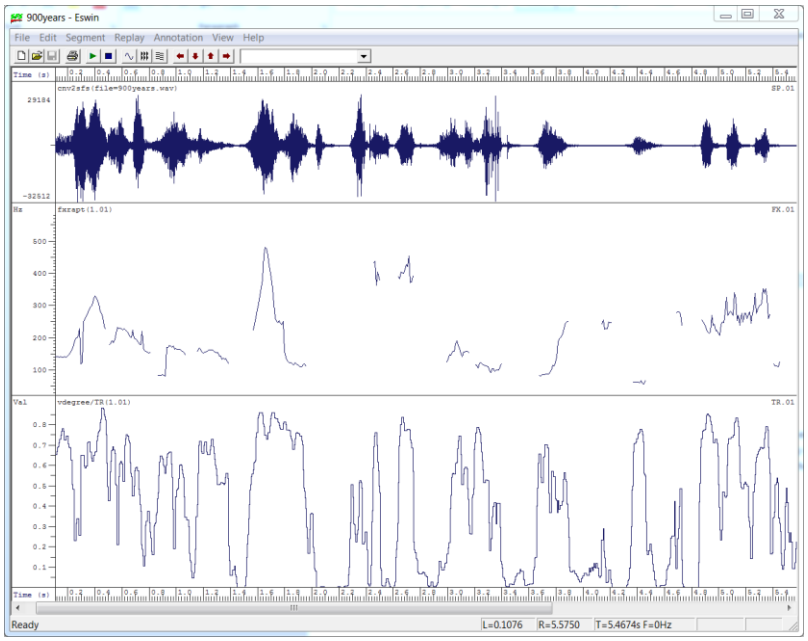


Figure 4 Screen capture of Speech Filing System showing waveform, f0 and degree of voicing.

Speech Feature Extraction

Selection Criteria

Paralinguistic speech features were selected for extraction and analysis based on the following criteria:

- Being reasonably believed to be a component of a listener's perceived *speaking style.*
- Meaningful to humans directly

Paralinguistic speech features from four difference classes are used in affect detection: prosodic, spectral-based, articulatory-based, and voice quality (see Figure 5). While they all may contain relevant information to affect and charisma, only prosodic features are directly meaningful to humans. Differences of pitch and speech rate between two speakers are readily understandable to humans, but measures of differing frequency spectrums or vocal tracts changes are not. For this reason, the features selected for this system are prosodic and not spectral or articulatory.

The 'meaningful to humans' requirement restricts features to those which people can perceive and control directly, and which are therefore suitable for describing the differences between speakers and speaking styles, and suitable as targets for intentional training.

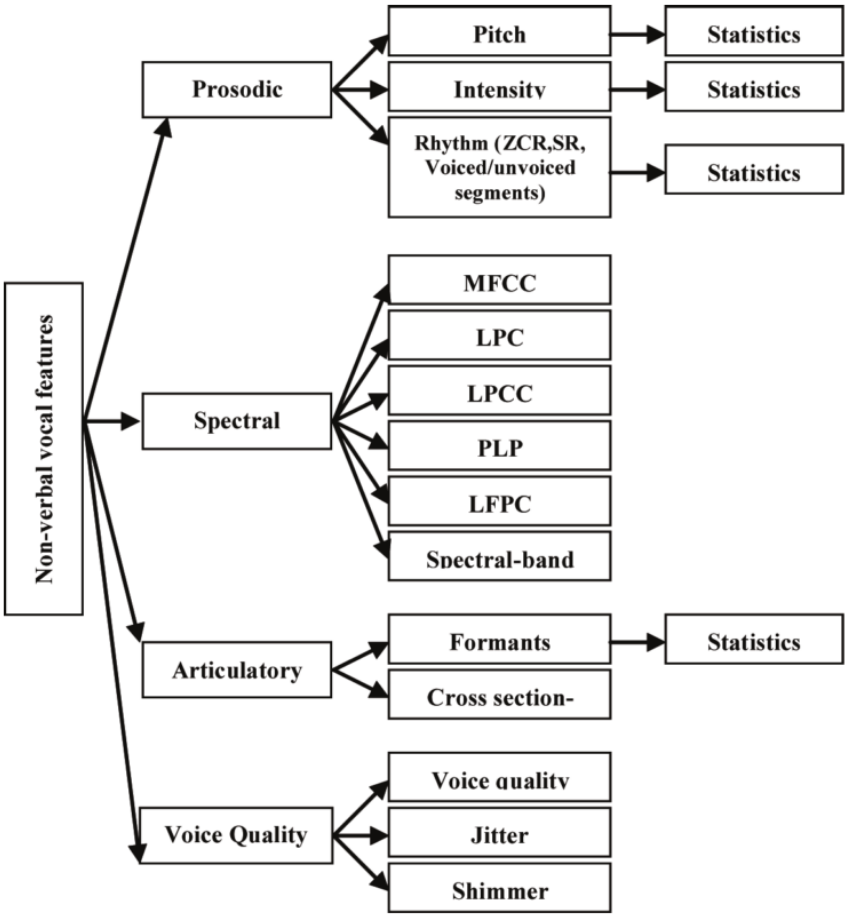


Figure 5 Feature Groups from Vinciarelli and Mohammadi, 2010 [3]

Pitch

What is Pitch?

Properly defined, pitch refers to the subjective auditory perception of tone by a person. For complex signals with many frequency components, perceived pitch is a nonlinear function of a signal's spectral and temporal energy distribution [12]. However, the fundamental frequency (f0) of a signal is an objective measure which correlates well with perceived pitch. The fundamental frequency is generally taken to be the inverse of the smallest true period in the period being analysed. In digital speech processing contexts, pitch and pitch tracking refers to f0, as it will here.

A person's pitch is core to how they sound when they talk, and is fundamental to their speaking style. People's typical pitch or mean pitch differentiates the sound of their voice from others, being deeper or louder. But more importantly, a person's variation in pitch and the patterns of their pitch variation dramatically determine how they come across. Contrast a person who speaks in a monotone with minimal pitch variation, an excitable person who has frequent rises of pitch, and a confident, determined person who ends their sentences with a commanding drop in pitch. As cited above, Rosenberg and Hirschberg, Strangert, and Strangert and Gustafson [6, 7, 8] found relationships between pitch and charisma. If the right pitch makes for charismatic speech, then the wrong pitch makes for uncharismatic speech, and it is clear that pitch is a key part of speaking style. It was thus selected as a feature to be analysed.

Mean, variance, minimum, maximum, and range of pitch values are global pitch statistics characterising a speaker. Additionally, the patterns of pitch variation (pitch contours) employed by a speaker provide further important detail of style; see below in Finality Patterns and Pitch Contours.

Calculating Pitch

There exists a considerable variety of approaches to calculating the pitch of signal. These include zero crossing rate, autocorrelation, cepstral, cross-correlation, and other methods. For this project, an accurate and easily available acronym was required. Talkin's RAPT (Robust Algorithm for Pitch Tracking) [12] is widely used and widely available. An evaluation performed by Gonzalez and Brookes [13] showed that for high SNR signals, RAPT performs equally well to newer more sophisticated methods (see Figure 6). The PEFAC algorithm [13] was explored, but it failed to resolve detail needed for pitch contour processing adequately. Accordingly, the implementation of the RAPT algorithm in SFS was used for estimating the pitch of speech recordings in this project.

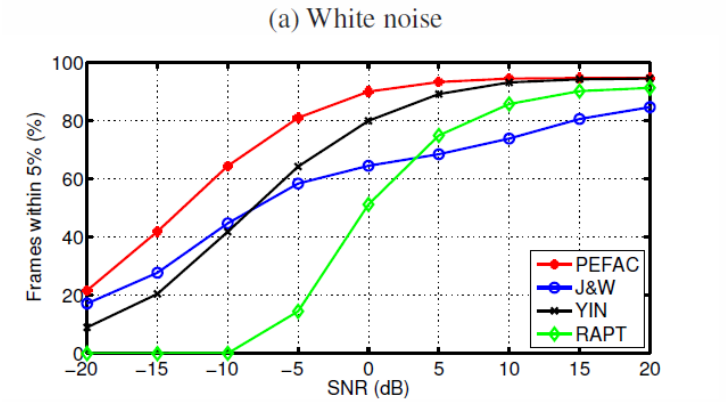


Figure 6 Comparison of Pitch Estimation Accuracy vs SNR for several algorithms, from Gonzalez and Brookes, 2011 [13] Shows that in high SNR signals, RAPT performs as well as other algorithms.

Challenges

Unfortunately, the extraction and use of pitch from a signal is not straightforward. The frame length for the pitch analysis must be decided, and further processing such as filtering and interpolation must be designed. Figure 7 compares the RAPT pitch contour in SFS with various forms of processing. Comparing the second and third panels, it can be seen that applying filtering to the pitch contour can remove crucial information – in this instance, masking the upward inflexion at the end of the utterances.

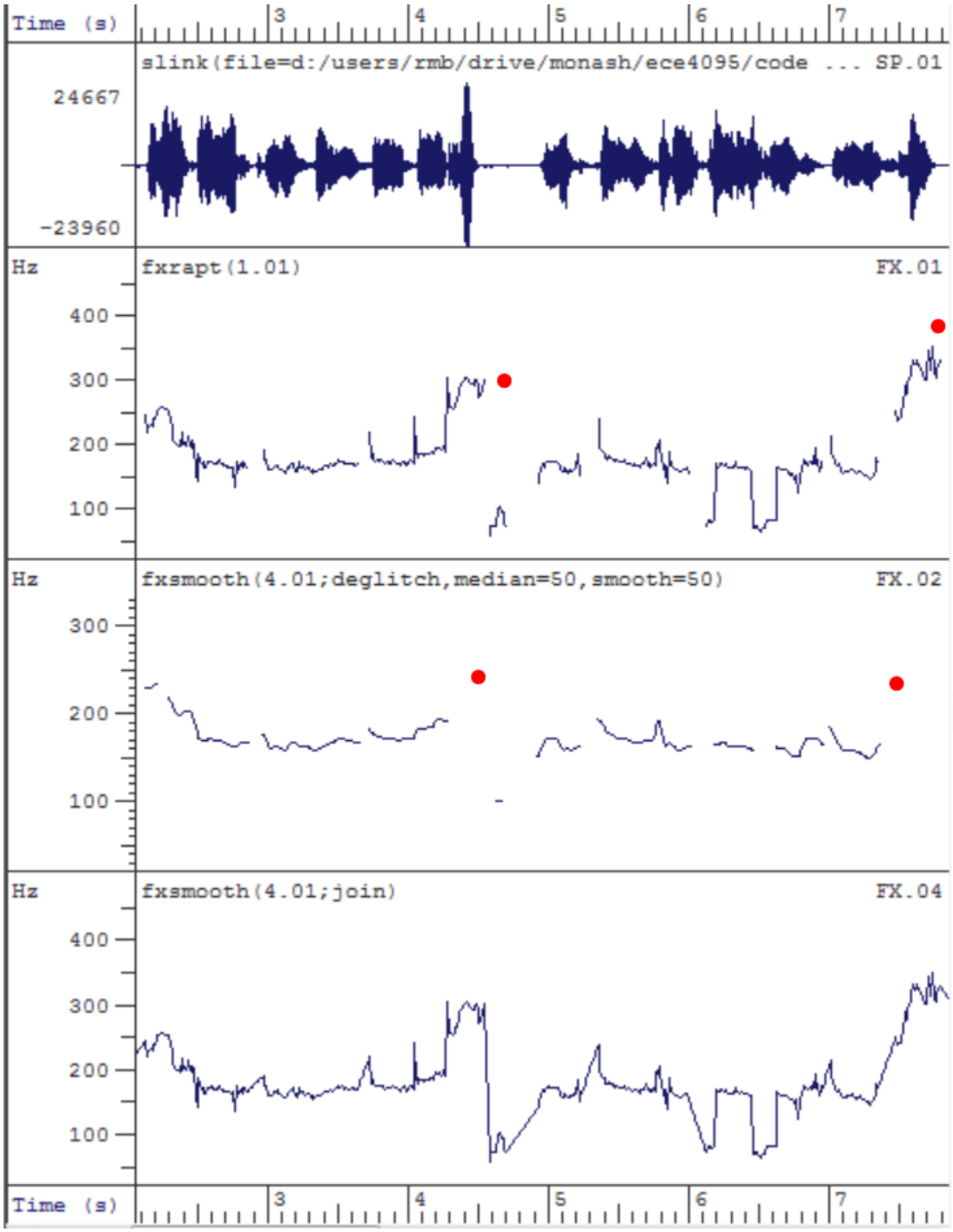


Figure 7 Speech Waveform and pitch contour with various processing in SFS. First panel: speech waveform. Second panel: unprocessed pitch contour. Third panel: pitch contour with deglitching, median filter and linear (cosine) filter applied. Fourth panel: pitch contour with unvoiced regions joined (interpolated). The red dots indicate where the filtering removes the prosodically significant upward inflexion at the end of utterances.

Voice Activity Detection

What is Voice Activity Detection?

Voice Activity Detection (VAD) is the detection of presence or absence of speech in a recording at any time point. Typically, VAD is performed on frames of speech signal, with each frame being marked as having or lacking speech. VAD is widespread in telecommunications where VAD can be used to save bandwidth by not transmitting unvoiced frames. Importantly for paralinguistic analysis, VAD can be used to identify the pauses in a speech recording, and conversely, the segments of speech constituting utterances (to be defined precisely below). Furthermore, VAD can be used to segment a speech recording into separate utterances, which can then be analysed on their own.

Pauses, utterances, and segmentation give VAD prime importance in this project. The pattern of pauses and utterances in a speaker's speech contribute enormously to the speaking style. Contrast a speaker who speaks without taking a breath with a speaker who has measured pauses between each statement for effect, or compare a speaker who pauses frequently in a jarring way out of disfluency and inability to recall what they were going to say next. The length of one's utterances is the dual of pauses and contributes in the same way. In this project, an utterance is defined prosodically as the dual of a pause and is a segment of uninterrupted speech between pauses; it is not a syntactic or semantic unit. See Figure 8 for the variation in speaker's Speech/Pause track.

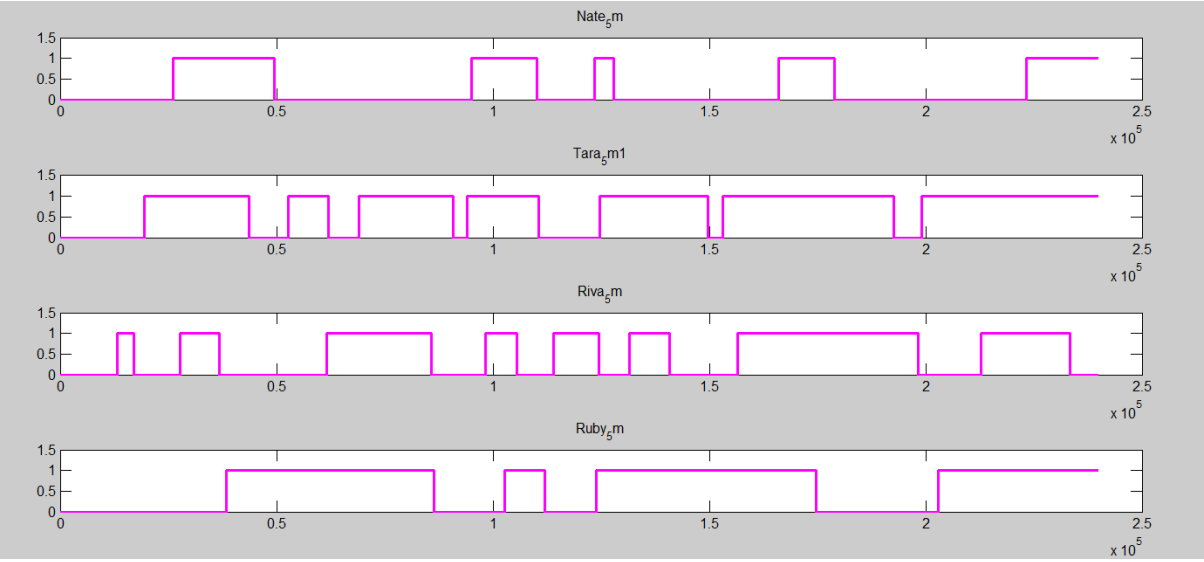


Figure 8 Speech/Pause track for several speakers. High for speech, Low for pause. The different patterns for each speaker are readily apparent.

Calculating VAD

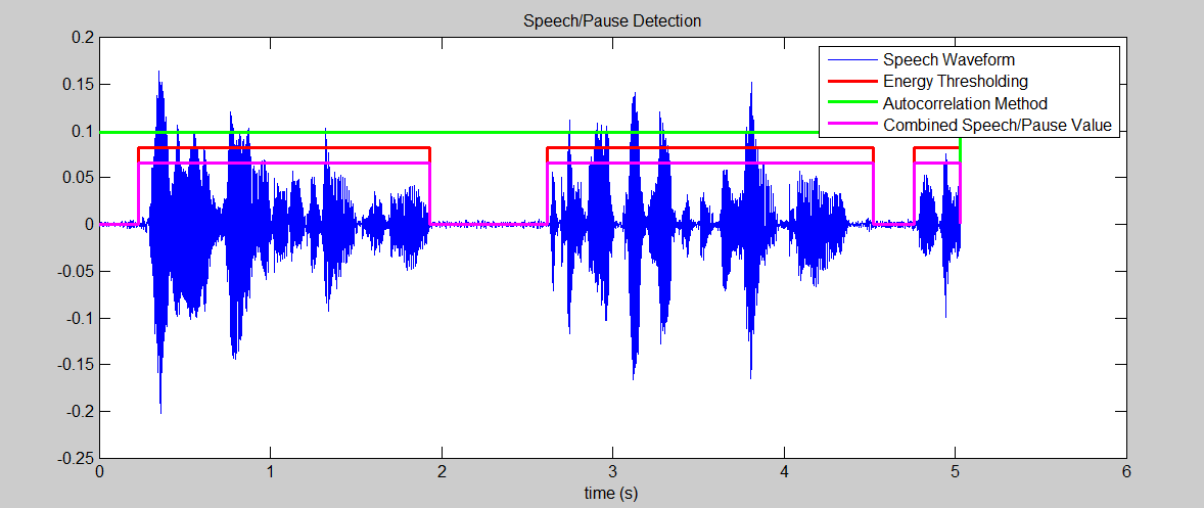
There exist many approaches to VAD: common algorithms use energy thresholding, Fourier coefficients, periodicity, zero-crossing rates, and increasingly, statistical models [14]. As VAD is of prime importance in this project, considerable effort was invested in attaining an algorithm with excellent performance.

Over the course of the project, three successive approaches to VAD were tried.

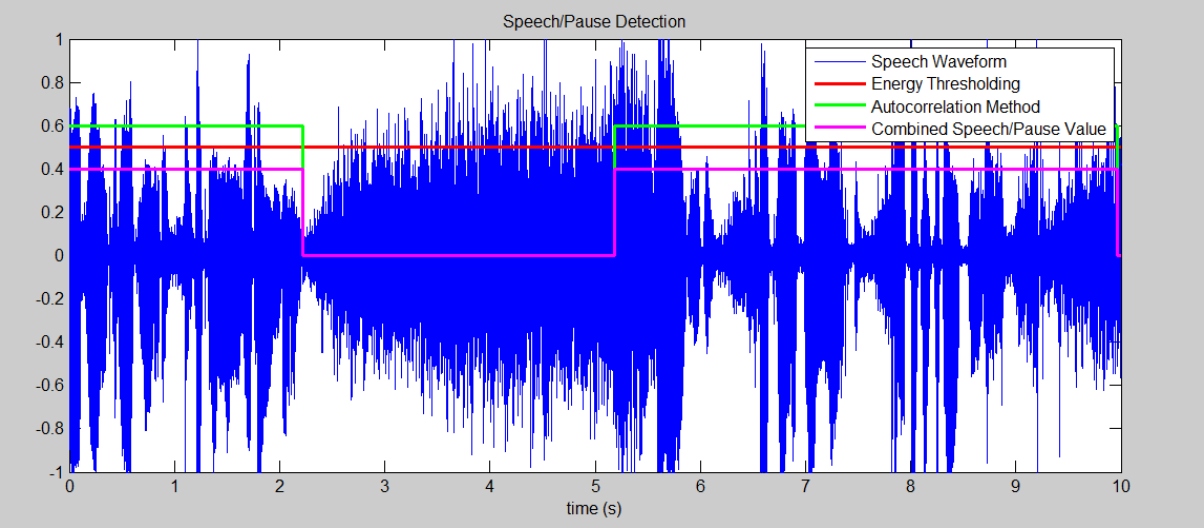
The first approach combined energy thresholding with autocorrelation detection. Energy thresholding detects the presence or absence of speech in a signal by comparing the energy in a frame with some reference value, in our case, the maximum energy of any frame in a signal. Frames

with an energy level 30dB less than the maximum energy in the signal are judged as non-speech. An autocorrelation approach seeks periodicity in the signal to distinguish voice, periodic parts of the signal, from non-periodic noise sections. A frame was deemed as containing speech if its autocorrelation had a peak which was 0.4 or greater of the lag 0 peak. The autocorrelation method was adapted from [15].

Energy thresholding works well for discriminating speech/non-speech in low noise signals where non-speech sections have low energy, but performs poorly with noisy signals. Conversely, the autocorrelation method works well in noisy signals where the high periodicity of spoken regions can be contrasted with the aperiodicity of noisy sections. Combining the output of two methods provides a versatile algorithm which works well for both noisy and non-noisy signals. See Figure 8 for comparison of these approaches on each signal type.



a)



b)

Figure 9 a) Output of Energy Thresholding method and Autocorrelation Method on a high SNR signal. The energy thresholding identifies the signal accurately while the autocorrelation method does not. The signals are combined by an AND operation. b) Output of Energy Thresholding method and Autocorrelation Method on a signal with significant noise in pauses. The autocorrelation methods detects the speech accurately and the energy thresholding does not.

The second approach to VAD attempted in this project was to use the pitch track generated by the SFS RAPT algorithm is an indicator of speech: any section of the signal with a detected pitch value

contains speech. Essentially, the pitch output of the RAPT algorithm operates in the same way as the autocorrelation method used in the first approach, but is more sophisticated and consequently more accurate. Unfortunately, VAD methods which rely on detecting periodicity in the signal can only detected *voiced frames*, but not *unvoiced frames*[3]. This is introduces some inaccuracy into these methods.

In pursuit of even better performance, a cutting-edge statistical method was employed. Tan and Lindberg [16] use a two-pass segment-based unsupervised method. Matlab source code for their algorithm was made available online, and this was successively integrated into the project. Visual inspection of VAD output (high and low) was plotted over the speech waveform, and playback of segments of speech identified by VAD confirmed that the Tan and Lindberg algorithm gave superior performance to previous methods.

Challenges

Identifying the 'pauses' in a speech recording is not as straightforward as considering any non-speech region output by the VAD as a pause. It is necessary to decide how long a pause must be before it count as a 'true pause'[4]. This decision needs to be made with a consideration of the purpose of the pause detection. Pause types of interest could be longer *rhetorical pauses* used to break up speech for delivery and impact (400+ms), or *syntactic pauses* which separate sentences (200-300ms). Based on my examination of speech recordings, a pause must be 400ms or longer to easily noticeable in speech. Additionally, it is very difficult to use pauses alone (or even an assemblage of prosodic features) to segment speech syntactically, as pauses within a sentence or clause can easily exceed pauses between sentences or clauses.

The dual problem of identifying pauses in a speech recording based on VAD output is that of using the VAD to segment the speech recording into utterances. Here an utterance is any section of speech recording between pauses of sufficient length, where sufficient length must be decided in a similar manner to minimum pause length. As stated above, in this project an utterance is not a semantic or syntactic unit, but a prosodic/acoustic one. However, it is desirable for the prosodic utterances to line up with syntactic units as much as possible. This is particularly important if prosodic analysis such as pitch contours and finality patterns (discussed later) are to be conducted on extracted utterances. For example, if the finality pattern of an utterance is to be calculated, it is highly desirable that the end of the utterance represents the true end of a spoken phrase.

For this project, it was decided that the minimum permissible pause duration would be 200ms, which was found to best segment periods of speech –utterances- into appropriate, meaningful units. The minimum utterance length was set at 300ms, however the later analysis of finality pattern was only performed on utterances of 1000ms or greater.

[3] These are phonetics terms where voiced speech are sounds produced involving the vibration of the vocal cords, and are consequently periodic. Unvoiced speech produced without vibration of the vocal cords lack periodicity.

[4] In fact, the ToBI (tone and break indices) annotation system distinguishes between four levels of break [63].

Finality Pattern

What is Finality Pattern?

Finality pattern refers to the pitch movement at the end of an utterance, e.g. whether pitch increases, decreases, or stays level (see Figure 10). While the overall level of pitch variation can discriminate a speaker with a monotone from a dynamic speaker, examining patterns of pitch variation such as finality patterns can further discriminate between speakers and speaking styles. Examples are clear after an introduction to the finality patterns typical in spoken English.

Broadly, English sentences will end with a drop in pitch for a declarative statement or command, and a rise in pitch for a question. This gives rise to the folk impression -with an arguable degree of linguistic truth- that speakers with frequent drops in their speech are assertive and confident, and those with frequent rises indicating questions are unsure and insecure. Given the bearing of finality patterns on how a speaker is perceived, they can be deemed a relevant feature to be included in speaking style.

High Rising Terminal

As stated, an upward inflexion at the end of a sentence in English typically indicates a question. Of late, considerable attention has been paid to the fact than many speakers and groups of speakers characteristically end in an upward inflexion even on declarative statements, e.g. introducing oneself with “My name is Brittany?” This tendency to conclude declarative sentences with a rise in pitch is called High Rising Terminal (HRT) or High Rising Intonation, and colloquially as upspeak or uptalk. HRT is considered a typifying feature of Australian English, and in a different form is a growing speech tendency in the US labelled “Valleyspeak”.

The attitude towards HRT is mixed. As above, many decry HRT as indicative of insecurity, uncertainty, and also submissiveness [17, 18]. Others have responded in defence of the phenomena, saying that it is a complicated linguistic phenomenon which does not indicate insecurity, but rather performs valuable linguistic functions such as floor-holding (indicating that one is not finished speaking) and keeping the listener engaged [19, 20].

It is not the goal of this project to take a prescriptive stance towards HRT, but it is hoped that the creation of tools for the automatic detection of HRT will facilitate further linguistic research into the phenomenon, and the creation of tools which grant users awareness of their own speech tendencies.

Falling Intonation

The dual of HRT is Falling Intonation (FI); this is the typical ending of declarative English sentences. Emphatic sentences or commands will often have a significantly marked drop in pitch and the end.

A: High Rising Terminal

B: Falling Intonation

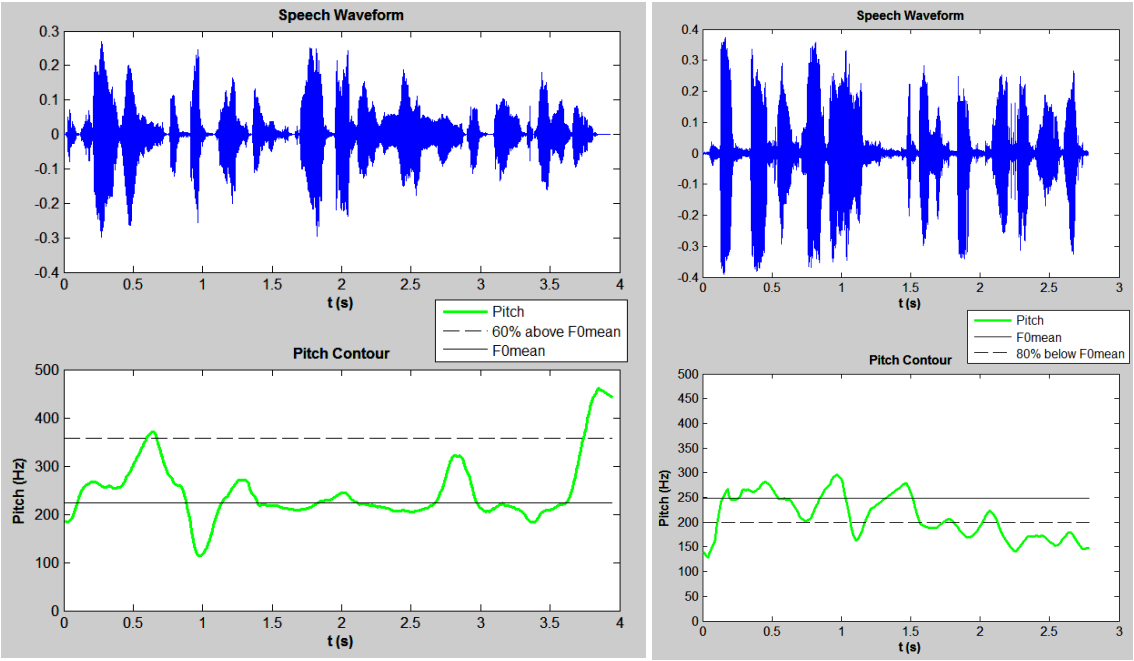


Figure 10 a) Example of an utterance with HRT. The final 0.5 seconds in the utterance demonstrate a pitch excursion to 450Hz, double the speaker's mean pitch and in excess of the 60% above mean threshold level for detecting HRTs. Utterance: "The children affected are described by those around them as being ↗ clumsy."
b) Example of an utterance with FI. The final 0.5 seconds in the utterance have a pitch value lower than both speaker's mean and lower than the rest of the utterance. Utterance: "But that is not enough to bring us back to our ideal ↘ weight."

Calculating Finality Pattern

The finality patterns in a speech recording can be extracted by segmenting the recording into utterances, calculating the pitch contour at the end of the utterance, and using a rule to classify the finality pattern.

The VAD algorithm is used to segment speech recordings into utterances, and those of 1s duration or longer have their finality pattern detected. As mentioned above, the utterances identified in this project do not represent a syntactic unit such as sentences or clauses, but a prosodic unit of uninterrupted speech. It is judged that this does not decrease, and may even increase, the validity of the measure. Consider the sentence "And I cut waste, increased efficiency, and boosted productivity." Classifying the finality pattern of the whole sentence will miss detecting rises at the end of each listed item as in common in users of HRT: "And I cut waste? . . . increased efficiency? . . . and boosted productivity?" The prosodic segmentation in this project would separate each listed item into an utterance on which it would classify the finality pattern.

For the selected utterances of 1s duration or longer, finality pattern was detected by two 'active' rules, one for detecting HRT and another FI. If the conditions for neither are met, the utterance has a flat finality pattern.

An utterance is classified as having a HRT if the peak pitch value in the last 500ms is 60% greater than the speaker's mean pitch. This value was selected by manually inspecting hand-labelled HRTs and deciding that 60% achieved the optimum balance between false positive and false negatives.

Asymmetrically, an utterance is classified as having a FI if 1) the mean pitch value of the final 500ms of the utterance were lower than the speaker's overall mean, 2) the mean pitch value of the final 500ms of the utterance were below a threshold percentage (80%) of the utterance mean pitch.

For each utterance, the ratios used for classifying HRTs and FIs were also recorded in addition to the decision values.

Challenges

The system devised for HRT and FI detection achieves excellent performance given its simplicity. Manual inspection of pitch contours and decision outcomes reveal almost no false negatives (see Figure 11 for graphical output used). However some false positives are evident. These arise from the limitations of the VAD based segmentation. Dynamic speakers will sometimes rise significantly midsentence and pause for a slightly longer duration than the minimum pause length used for segment utterance. Consequently, the mid-sentence rise is considered the terminal of an utterance and hence a HRT. Improved "intelligent" segmentation would likely alleviate this problem.

Figure 11: Graphical Output used for inspecting Segmentation, Pitch Contours, and Finality Pattern classification. The numbering of each utterance allows them to be easily identified and played.

Speech Rate

Speech rate is a feature was explored, but not developed to an adequate level of performance to be included in the final system or analysis. This was due to time constraints. In my judgment, speech rate is an important aspect of speaking style.

Detection of speech rate (syllables/second) was explored using the code and algorithm from Wang and Narayanan [21]. Their algorithm uses temporal and selected sub-band correlation to detect syllable nuclei – the nucleus of a syllable, typically the vowel. This method detects speech rate without using automatic speech recognition.

Fundamentally, syllable nuclei are detected by searching for energy peaks in the waveform surrounded by minima. A plot of detected peaks and minima is show in Figure 3. The number of these peaks is counted divided by length of the speech sample to give speech rate.

It was found that the algorithm tended to undercount syllable nuclei and underestimate the speech rate. Due to time constraints, this could not be explored further.

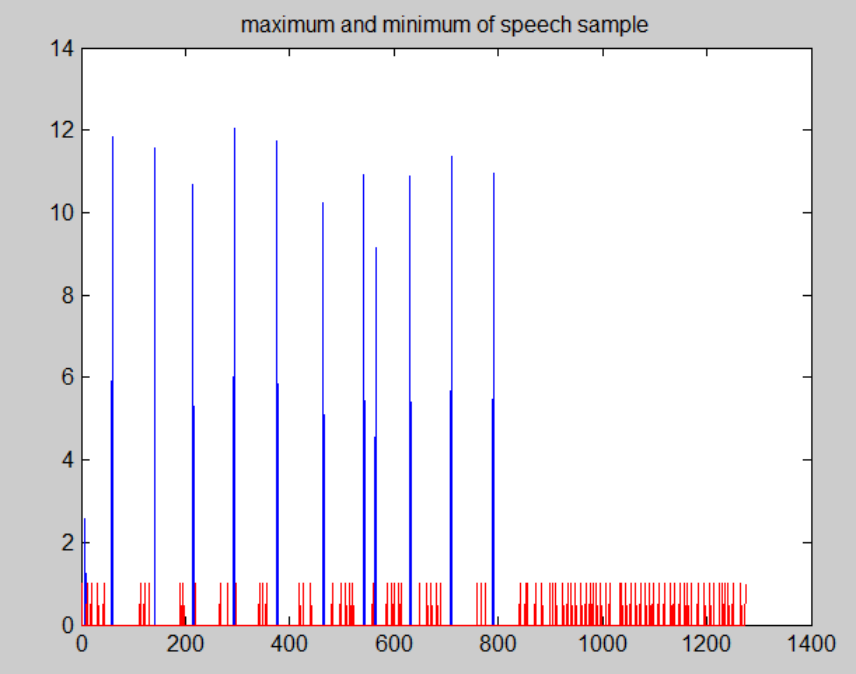


Figure 12 Peaks and minima of speech waveform energy showing syllable nuclei

Speech Type and Context

Speech can be classified in a large variety of types and contexts. These include read, rehearsed, spontaneous, formal, informal, conversation, interview, and many others. For this project, it was necessary to decide which kind of speech should be investigated.

Largely, two options are available for making interesting comparisons. First, speech recordings from one speaker in multiple contexts could be compared to discover the differences in speaking style in different contexts of a single speaker. This approach highlights inter-context differences. Cullen, Hines, and Harte [22] have constructed a database suitable for this purpose. Second, recordings from multiple speakers in the same or similar context could be compared to highlight inter-speaker differences.

For this project, it was decided to compare multiple speakers in the same context of formal presentation. Formal presentation here refers to an organised presentation or speech to an audience. This context was chosen because of people's generally high interest in improving their 'public speaking skills' and because people are often prepared to invest time rehearsing speeches and presentation. It is therefore assumed that information about what kind of public speaking styles are best received and tools for improving public speaking would be the most welcome. This makes formal presentation an excellent first target for paralinguistic analysis.

Speech Corpora

Developing a speech-based system requires appropriate bodies of speech recordings for development, testing, and analysis. For this project, suitable audio sources were sought for 1) use in development of code for extracting paralinguistic speech features, and 2) to test the relevance of the extracted features for capturing speaking style.

Labelled and annotated speech corpora generally serve as the ground truth to testing the performance of speech-related algorithms. The Aix-Marsec database [23] is a heavily-annotated, machine readable speech corpus derived from successive generations of annotation on the original Spoken English Corpus. Figure X shows a sample of speech from the database with annotations in the Praat Text Grid system.

Secondly, the SSPNet Speaker Personality Corpus [24] was obtained for the relevance of its content for how speakers are perceived. The Speaker Personality Corpus contains clips from subjects rated on the Big Five Personality traits.

Regrettably, due to limited time, extensive testing of the system was not conducted with these corpora.

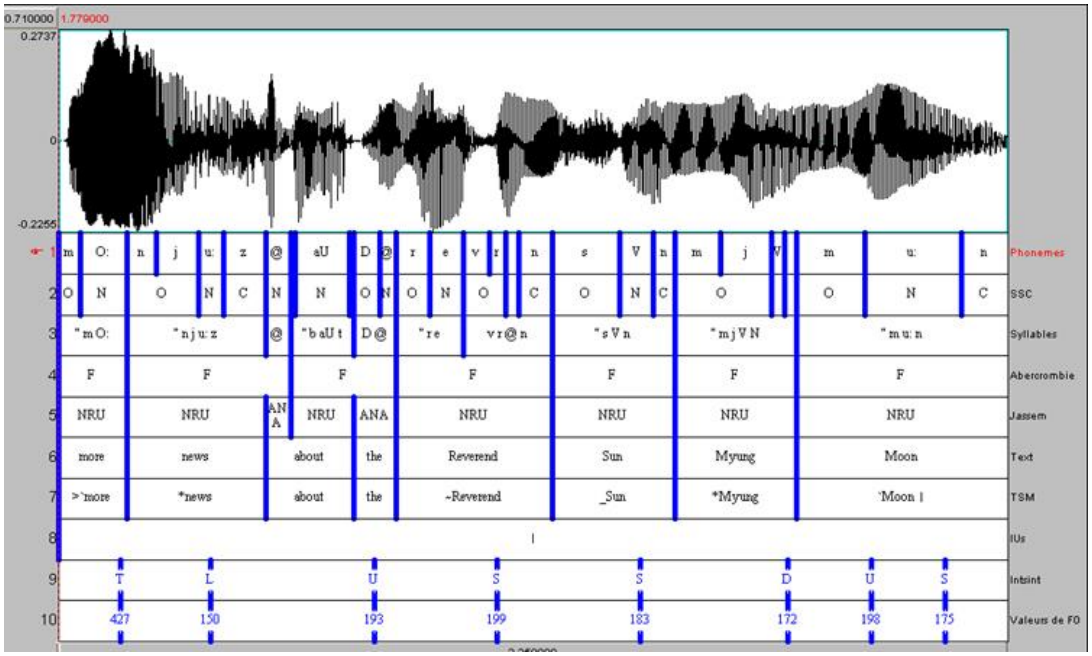


Figure 13 Prosodically Annotated Speech Sample in Praat Text Grid, showing phonemes, syllables, and text.

Additionally, neither of the above two corpora contained speech recordings of the desired type for speaking style analysis. The combination of speech corpora being difficult to obtain and lack of any easily findable speech corpora containing formal presentations meant it was decided to self-source collections of audio recordings. Collections of self-sourced recordings have the benefit that they can be more precisely chosen for the purposes of the project.

Speech recordings from three groups of speakers were collected. Differences between speakers and between the groups of speakers are of interest (see Table 1).

Group	N	Description	(Male/Female)	Assumed Charisma
Wedding Speakers	6	Ceremonial speeches at wedding.	3m, 3f	Medium
Student Oral Presentations	7	Group presentations to the class on the topic of a genetic disorder.	7f	Low
TED Speaker	7	Presentations given to a large audience and recorded and made available online on a high traffic website. Each talk received a minimum of 1 million views.	7f	High

Table 1 Three groups of speakers analysed using the system.

Group 1: Wedding Ceremony Speakers

The wedding ceremony of the author consisted of several formal, ceremonial speeches regarding the commitment of the wedded to each other and to their values. The tone of the speeches was overall solemn and reverential. The speakers were articulate and charismatic, although not to professional levels.

Group 2: Student Oral Presentation

As a required part of their course, 2nd year Monash psychology students are required to deliver a group presentation on a given genetic disorder. These recordings were used with permission from Blackboard recordings of the presentations.

Group 3: TED Speakers

Seven TED talks each receiving a minimum of 1 million views on the website were sourced. The speakers were selected to match the Student Oral Presentation group, with all speakers being female and talks on topics related to psychology being preferred. Due to the nature of speeches and the large audience, it is assumed that this group of speakers possesses the greatest level of charisma.

Each group is a distinct population speaking a similar, although not identical, situation. Listening to the speech recordings, it is readily apparent that the speakers individually and the group overall differ in their speaking style. Hence, if the paralinguistic speech feature extraction works accurately and the features were selected appropriately, then the system developed should reveal differences between speakers and groups of speakers.

The recording conditions for each group were not controlled and may compromise results.

Results and Discussion

Intro

As described above, the goal of this project is three-fold.

1. To develop and test tools for extracting paralinguistic speech features from speech recordings.
2. To identify speech features which individually or collectively contribute to a speaker's distinctive speaking style.
3. To test the selection and extraction of speech features on a collection of speech recordings to determine whether the selected features and the method of extracting captures differences between speakers and groups of speakers.
4. To potentially identify differences between groups of speakers.

Regarding 3., it is not the goal to apply the developed tools to samples of speakers and infer differ characteristics of the populations. Instead, we start with populations we know to be distinct and expect to vary in specific manners. The system developed is successful if the system can automatically capture differences which are apparent to humans. For example, a typical human who is paying attention can easily identify that TED talkers speak slower, have longer pauses, greater number of emphatic statements, and fewer HRTs than do undergraduate psychology students. If the system indicates differences between these two speaker populations, we can infer that the system is capable of capturing existing differences.

Additionally, while a human can easily notice that one speaker tends to raise their pitch more than another, the system can add quantification and precision on to human skill. Moreover, it can be used to extract many features simultaneous, while a human will become overtaxed if they try to attend to too many aspects of speech at once.

Regrettably, owing to the limited sample sizes it is not possible to make inferences about the different groups with statistical validity. However, to the extent that findings of differences between the groups are in the expected direction, this is some measure of evidence of favour of the tools developed being useful.

Group Differences

Listed below are notable differences between the groups; see also Table 2. It must be noted that the speakers in Student Presentation group and TED Speakers group are all female, whereas the Wedding Speaker group is half female, half male. The male voices have significantly lower pitch and accompanying lower pitch variation, which affect the values for this group. Additionally, given that speaking style generally differs with gender, the focus will be on comparisons between the Student Presentation group and the TED Speaker group.

- TED Speakers had both higher mean pitch values and pitch variation than the other groups. This is expected given the findings of [6, 7, 8] regarding pitch and charisma, and the expectation that the TED speakers are the most charismatic group.
- TED Speakers have pauses of slightly longer duration than Student Presentation group; standard deviation of pause length is similar.
- Student Presenters have longer utterances on mean, which is equivalent to longer stretches of speech without pause, however the variation in their pause length is considerable.
- Student Presenters spoke for 80% of the time of speech recordings compared with 76% for TED Speakers on average.
- 16% of utterances from Student Presenters concluded with HRT on average, compared to 6% for the TED Speaker group and 0% for the Wedding Speaker Group.
- The TED Talker group concluded 26% of sentences with FI on average, compared with 10% of the Student Presenters and 15% for Wedding Speakers.

  In total, the system found that TED Talkers varied their pitch more, paused more often and longer, ended their sentences emphatically, and did not use HRT compared to Student Presentations.

	Student Presentations	TED Speakers	Wedding Speakers
Length (s)	261	300	300
f0 Mean (Hz)	178	222	166
f0 Std (Hz)	43	53	30
Mean Pause Length (s)	0.689	0.745	0.980
Pause Length Std (s)	0.462	0.432	0.702
Mean Utterance Length (s)	2.837	2.491	2.348
Utterance Length Std (s)	5.325	3.779	1.951
Speech/Pause Percentage (%/%)	80/20	76/24	71/29
HRT Percentage (%/100)	0.16	0.06	0.00
FI Percentage (%/100)	0.10	0.26	0.15

Table 2 Mean of each speaker group for each variable listed.

Individual Differences

Equally interesting to the group differences are the extreme range of differences captured between individuals (n_{total} = 20). Table 3 compares the range of values of the top 5 scoring and bottom 5 scoring individuals for each measure. See Appendix A for complete results for each speaker.

	Range of Top 5	Range of Bottom 5
Mean Pitch	220-260Hz	155-161Hz[1]
Pitch Standard Deviation[2]	73-46Hz	30-42Hz[1]
Mean Pause Duration	900-1600ms	500-600ms
Pause Duration Std[3,4]	900-1500ms	100-160ms
Pause Percentage	28-42%	11-18%
HRT Percentage	13-42%	0-0%
FI Percentage	20-62%	2-5%

Table 3 Range of scores for highest scoring and lowest scoring five individuals on each measure.

1. These are scores the lowest scoring females. The three males scored lower than all females and had mean pitch values between 117-145Hz, and pitch standard deviation values of 18-30Hz.
2. Mean Pitch and Pitch Stand Deviation are correlated at R = 0.78, $p < 0.0001$, 95% CI [0.52 .91].
3. Events during recording mean that these values may be affected by outliers.
4. Mean Pause Duration and Pause Stand Deviation are correlated at R = 0.48, $p < 0.05$, 95% CI [0.05 0.76].

Speakers differ dramatically on each of the measures. This indicates that the features extracted are capturing at least some meaningful differences between speakers which constitute their speaking style.

However, many of these differences cut across group boundaries. This suggests that while the measures do differentiate speakers, if we accept that some groups are more charismatic than others, than the features extracted are not identifying the measures which correlated strongly to charisma. In other words, it may be possible to be charismatic with great or little pitch variation, a fast speaking rate and few pauses, or slow speaking rate and long pauses.

Consider that the system at present does not detect whether a lengthy pause between statements is a deliberately timed pause for rhetorical effect or simply the result of disfluency. Additionally, there are higher order measures such as the overall rhythm and pattern of pauses which have not yet been measured. In support of the notion that it requires higher order measures to capture differences of charisma, the measure of HRT and FI – higher order measures compared to simple mean pitch and pitch variation – did reveal large differences between the groups.

Conclusions

In summary, the most significant outcomes of this project are:

Selection of Features and Methods of Their Extraction

- Pitch and pitch statistics, pause/utterance and pause/utterance statistics, finality patterns (High Rising Terminal and Falling Intonation) were identified and verified as paralinguistic speech features which contribute significantly to a speaker's distinctive speaking style.
- Code for the automatic analysis of the above speech features was developed.

Application of the Developed Tools to Test Groups of Speakers

- The speaker groups differ considerably on their mean pitch, percentage of utterances ending in HRT, and percentage of utterances ending in FI.
- The differences between individuals was even starker. For all measures, the range of values for top 5 scoring and lowest 5 scoring individual did not overlap.
- The differences between individuals matches easily noticeable differences in how their recordings sound.
- Regrettably, limited sample sizes mean that valid inferences cannot be made from the group analysed. The outcomes in the expected direction merely offer some degree of confidence in the tools developed.

Limitations

A number of limitations affect this project and should be addressed in further work.

- Some inaccuracies in the segmentation into utterances have downstream effects in allowing some misclassification of HRTs and FIs.
- The small sample sizes used in the analysis undermine statistical validity.
- The lack of rigorous testing of feature extraction methods.
- Lack of control of recording environments for the audio analysed may be introduce confounds which were not controlled for.

Outcomes

The overall goal of this project was to use computational paralinguistic methods to develop tools which would be useful in the scientific and engineering analysis of speech, speaking style, and charisma. This project can be deemed a success. It succeeded in achieving the following targets:

- Identifying paralinguistic speech features which contribute to a speaker's perceived speaking style.
- Developing tools for the automatic extraction of these features from speech with low complexity and low computational cost.
- The developed tools succeed in quantitatively identifying the differences between speakers.
- The developed tools were used for a test analysis demonstrating their practical usefulness.

It must be noted that the requirements of the project changed midway due to the change from old to new supervisor. Modified requirements were not specified. In the original project design, one of the major project aims was to develop an end-user system for analysing speech. This aim was changed to instead focus on the development of the tools and application of them to a set of test speech recordings in order to verify them. Notwithstanding, most of the original formal requirements were met and they are listed below together with outcome status.

The complete requirements analysis document is included in Appendix B.

[R.001] The system will be accompanied by a literature review summarising research and techniques relevant to the system to be developed.

> ✓ The background section in this document provides research relevant to development of the system.

[R.002] The system must analyse a speech sample from a user and extract at least four paralinguistic speech features such as pitch, speech rate, and number of pauses. The combination of these results constitute the speaking style in the given sample.

> ✓ The system successively extracts pitch, pauses, utterances, and finality patterns.

[R.011] The system will work with any recording of a quality equivalent to at least 16bits per sample, 16,000Hz sample rate and of reasonable SNR (to be determined empirically).

> ✓The system works across a broad range of audio recording qualities. It has been tested extensively with 16,000Hz of varying SNR.

[R.012] The speech sample may be a recording of a user speaking a predetermined passage.

> ✓ The system would work with a predetermined passage, but as per OR.012, it works with any speech sample.

[OR.012] Optionally, if a more advanced system is feasible than the speech sample used may be any short speech sample from a user.

> ✓ The system works with any recording of single speaker speaking.

[R.003] The selected speech features must have evidence of being involved in human perception of speaking style. Evidence can be sourced from research in affective computing, relevant social sciences, voice coaching practices, speech pathology practices, or empirical research.

> ✓ The features extracted are supported as being relevant to speaking style by the literature reviewed in the Background section, any by their success in differentiating speakers

[R.004] Each speech feature measurement must be checked for validity against an external measure, e.g. manual transcription of speech rate or pauses, comparing pitch analysis with alternative algorithms.

> ✗ The feature measurements were not thoroughly checked for validity.

[R.005] The system must be able to reliably indicate a difference in the speaking styles of two speech samples if humans rate the samples as having significantly different speaking styles.

> ✗ This capability was not tested.

[OR.005] If there are cases in which the features measured in the system fail to reflect differences in speaking style which humans consistently detect, an investigation will be conducted.

> ✗ The failure of system to identify greater differences in the pause behaviour of TED Speakers and Undergraduate speakers was considered.

[R.006] The system will be accompanied by an explanation of the quantities measured which is meaningful to the user. For instance, an explanation of what pitch is and it is relevance to perceived speaking style.

> SUPERCEDED: This requirement was relevant under the original project aims of developing an end-user product.

[R.007] The results of the analysis must be displayed graphically.

> SUPERCEDED: This requirement was relevant under the original project aims of developing an end-user product. However, graphical output was developed for examining the output of analyses.

[R.008] The user interface of the system must allow the user to compare at minimum two different speaker profiles.

> SUPERCEDED/✓: The current project does not include a user interface, however the system allows for comparison of multiple speakers.

[R.009] The analysis will take no longer than 120s for a 30s speech sample on a PC with 1GHz or faster processor, 1GB RAM, and 16GB available hard disk space.

> ✗ Untested: The system processes speech faster than real-time on a powerful PC.

[OR.009] The processing may be moved to a server to meet real time requirements.

> SUPERCEDED: This is not relevant to the non-end user system.

[R.010] The end system will be an application usable on a widely available device type and operating system.

SUPERCEDED: This is not relevant to the non-end user system.

[DG.010] The end system may take the form of a stand-alone Windows application, compatible with Windows 7 32-bit OR a web-application designed to be usable by a browser with either HTML5, Adobe Flash Support, or WebRTC depending on implementation. If a web-application, the system will be guaranteed to work successfully on at least one web browser. OR, the system may be implemented on the Android smartphone operating system with either local or server processing.

SUPERCEDED: This is not relevant to the non-end user system.

[C.001] The system will be designed and tested with native English speakers. Full functionality and reliability are not guaranteed with speech samples from other languages or from non-native English speakers.

The system was only tested on speech from native English speakers with a single exception.

Future Directions

While much excellent work has been completed, many refinements to current work and further extensions are possible. Future work should include:

- Rigorous testing of extracted measures against ground truth sources.
- The addition of additional speech features, e.g.
    - Speech rate
    - Filled pauses (umms and ahhs)
    - Energy variation.
    - Pitch Contour analysis over entire utterances in addition to finality pattern.
- More sophisticated analysis of pause/speech patterns to further differentiate speakers; possibly a “frequency analysis” of the voice activity track.
- Improved speech segmentation, possible achieved via inclusion of Automatic Speech Recognition.
- Expansion of the work to different speech contexts, e.g. interviews, political speeches, comedy routines.
- Inclusion of speaker diarisation, the ability of a system to separate a stream of audio into separate speakers, so that speech recordings with multiple speakers such as conversations or interview can be analysed.
- Principal Component Analysis of extracted features and their statistics in order identify which elements are core to speaking style.
- Improved methods of HRT and FI classification which use machine learning techniques instead of simple rules.
- Collaboration with linguists.
- Combination of extracted paralinguistic features with automatically extracted linguistic features via automatic speech recognition for a comprehensive speaker profile.

Bibliography

[1] A. Mokharti and N. Campbell, "Speaking Style Variation and Speaker Personality," *Proc. of Speech Prosody,* 2008.

[2] Z. Zeng, M. Pantic, G. I. Roisman and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Tansactions on Pattern Analysis and Machine Intelligence,* vol. 31, no. 1, pp. 39-58, 2009.

[3] G. Mohammadi and A. Vinciarelli, "Towards a technology of nonverbal communication: vocal behavior in social and affective phenomena," in *Affective Computing and Interaction: Psychological, Cognitive and Neuroscientific Perspectives*, 2010, pp. 133-156.

[4] B. Schuller, S. Steidlb, A. Batliner, F. Burkhardtd, L. Devillersa, C. Müllere and S. Narayananf, "Paralinguistics in Speech and Language--State-of-the-Art and the Challenge," *Computer, Speech, and Language,* pp. 4-39, 2013.

[5] G. Mohammadi and A. Vinciarelli, "Automatic Personality Perception: Prediction of Trait Attribution Based on Prosodic Features," *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING,* vol. 3, no. 3, pp. 273-284, 2012.

[6] A. Rosengburg and J. Hirschburg, "Acoustic/Prosodic and Lexical Correlates of Charismatic Speech," 2005.

[7] E. Stranger, "What makes a good speaker? Subjective ratings and acoustic measurements," in *Proceedings from Fonetik 2007: speech, music and hearing, quarterly progress and status report*, 2007.

[8] E. Strangert and J. Gustafson, "What Makes a Good Speaker? Subject Ratings, Acoustic Measurements and Perceptual Evaluations," *INTERSPEECH,* vol. 8, pp. 1688-1691, 2008.

[9] [Online]. Available: http://www.personal.rdg.ac.uk/~llsroach/phon2/freespeech.htm. [Accessed 4 May 2014].

[10] M. Huckvale, "Speech Filing System. Tools for speech research," 2014. [Online]. Available: http://www.phon.ucl.ac.uk/resource/sfs/.

[11] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2014. [Online]. Available: http://www.praat.org/.

[12] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis,* vol. 495, p. 518, 1995.

[13] S. Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," *Signal Processing Conference, 2011 19th European,* pp. 451-455, 2011.

[14] J. Kola, C. Espy-Wilson and T. Pruthi, "Voice activity detection," *MERIT BIEN,* pp. 1-6, 2011.

[15] L. Rabiner, R. Schafer, K. Vedula and S. Yedithi, “Autocorrelation Pitch Detector,” 29 May 2015. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/45309-autocorrelation-pitch-detector. [Accessed 1 June 2015].

[16] Z. H. Tan and B. Lindberg, “Low-complexity variable frame rate analysis for speech recognition and voice activity detection,” *IEEE Journal of Selected Topics in Signal Processing,* vol. 4, no. 5, pp. 798-807, 2010.

[17] H. Davis, “The Uptalk Epidemic,” Psychology Today, 6 October 2010. [Online]. Available: https://www.psychologytoday.com/blog/caveman-logic/201010/the-uptalk-epidemic. [Accessed 1 June 2015].

[18] “Want a promotion? Don't speak like an AUSSIE: Rising in pitch at the end of sentences make you sound 'insecure',” Daily Mail Australia, 14 January 2014. [Online]. Available: http://www.dailymail.co.uk/sciencetech/article-2538554/Want-promotion-Dont-speak-like-AUSSIE-Rising-pitch-end-sentences-make-sound-insecure.html. [Accessed 1 June 2015].

[19] J. Hoffman, “Overturning the Myth of Valley Girl Speak,” The New York Times, 23 December 2013. [Online]. Available: http://well.blogs.nytimes.com/2013/12/23/overturning-the-myth-of-valley-girl-speak/?_r=1. [Accessed 1 June 2015].

[20] M. Seitz-Brown, “Young Women Shouldn’t Have to Talk Like Men to Be Taken Seriously,” Slate, 16 December 2014. [Online]. Available: http://www.slate.com/blogs/lexicon_valley/2014/12/16/uptalk_is_okay_young_women_shouldn_t_have_to_talk_like_men_to_be_taken_seriously.html. [Accessed 1 June 2015].

[21] D. Wang and S. Narayanan, “Speech rate estimation via temporal correlation and selected sub-band correlation,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005.

[22] A. Cullen, A. Hines and N. Harte, “Building a Database of Political Speech: Does Culture Matter in Charisma Annotations?,” in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge.*, New York, NY, 2014.

[23] G. Mohammadi and A. Vinciarelli, “Automatic Personality Perception: Prediction of Trait Attribution Based on Prosodic Features,” *IEEE Transactions on Affective Computing,* vol. 3, no. 3, pp. 273-284, 2012.

[24] C. Auran, C. Bouzon and D. Hirst, “The Aix-MARSEC project: an evolutive database of spoken British English,” in *Speech Prosody 2004, International Conference*, 2004.

[25] D. Braga and M. A. Marques, “The pragmatics of prosodic features in the political debate,” in *Speech Prosody 2004, International Conference*, 2004.

[26] R. Hincks, “Computer support for learners of spoken English,” *Diss. Speech and Music Communication, KTH,* 2005.

[27] E. Strangert, “What makes a good speaker? Subjective ratings and acoustic measurements,” in *Proceedings from Fonetik 2007: speech, music and hearing, quarterly progress and status report*, 2007.

[28] J. B. Hirschberg and A. Rosenberg, “Acoustic / Prosodic and Lexical Correlates of Charismatic Speech,” in *Proceedings of Eurospeech'05*, Lisbon, Portugal, 2005.

[29] S. G. Koolagudi and K. S. Rao, “Emotion recognition from speech: a review,” *International Journal of Speech Technology,* pp. 99-117, 2012.

[30] B. Schullera, S. Steidlb, A. Batliner, F. Burkhardtd, L. Devillersa, C. Müllere and S. Narayananf, “Paralinguistics in Speech and Language--State-of-the-Art and the Challenge,” *Computer, Speech, and Language,* pp. 4-39, 2013.

[31] M. Nakamura, K. Iwano and S. Furui, “Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance,” *Computer Speech & Language,* vol. 22, no. 2, pp. 171-184, 2008.

[32] A. Veiga, D. Celorico, J. Proença, S. Candeias and F. Perdigão, “Veiga, A., Celorico, D., Proença, J., Candeias, S., & Perdigão, F. (2012). Prosodic and Phonetic Features for Speaking Styles Classification and Detection,” *In Advances in Speech and Language Technologies for Iberian Languages,* vol. 2012, pp. 89-98.

[33] O. F. Cabane, The Charisma Myth: How Anyone Can Master the Art and Science of Personal Magnetism, Portfolio Trade, 2013.

[34] M. Nakamura, K. Iwano and S. Furui, “Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance,” *Computer Speech and Language,* vol. 22, no. 2, pp. 171-184, 2008.

[35] L. R. Rabiner and R. W. Schafer, Introduction to Digital Speech Processing, Now Publishers Inc, 2007.

[36] K. Runge and D. Lummer, “Study on the environmental impact of high voltage overhead lines and underground cables (380kV),” 13 February 2013. [Online]. Available: http://renewables-grid.eu/fileadmin/user_upload/Files_RGI/Karsten_Runge_Dennis_Lummer_OECOS_Environmental_Impacts_of_Undergrounding_Highest_Voltage_Transmission_Lines.pdf.

[37] E. Shriberg, A. Stolcke and D. Jurafsky, “Can prosody aid the automatic classification of dialog acts in conversational speech?,” *Language and speech,* vol. 41, no. 3-4, pp. 443-492, 1998.

[38] S. Sudhoff, D. Lenertova, R. Meyer and e. al, Methods in empirical prosody research. Vol. 3., Walter de Gruyter, 2006.

[39] W. Wang, A. Stolcke and J. Yuan, “A Cross-language Study on Automatic Speech Disfluency Detection,” *HLT-NAACL,* pp. 703-708, 2013.

[40] A. Batliner, A. Kießling, S. Burger and e. al, “Filled pauses in spontaneous speech,” SciDok, 2011.

[41] M. Bhargava and T. Polzehl, “Improving automatic emotion recognition from speech using rhythm and temporal feature,” in *ICECIT 2012, Srinivasa Ramanujan Institute of Technology*, Anantapur, 2013.

[42] N. H. De Jong and T. Wempe, “Praat script to detect syllable nuclei and measure speech rate automatically,” *Behavior research methods.,* vol. 41, no. 2, pp. 385-390, 2009.

[43] T. Dekens, M. Demol and W. Verhelst, “A comparative study of speech rate estimation techniques,” *INTERSPEECH,* pp. 510-513, 2007.

[44] F. D'Errico, R. Signorello, D. Demolin and e. al, “The perception of charisma from voice: A cross-cultural study,” in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction.*, Geneva, 2013.

[45] J. Edlund and M. Heldner, “/nailon/-online analysis of prosody,” *Working Papers in Linguistics,* vol. 52, pp. 37-40, 2009.

[46] F. Eyben, F. Weninger and E. Marchi, “Likability of human voices: A feature analysis and a neural network regression approach to automatic likability estimation,” in *14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS),*, Paris, 2013.

[47] J. R. Green, D. R. Beukelman and L. J. Ball, “Algorithmic Estimation of Pauses in Extended Speech Samples,” *Journal of medical speech-language pathology.,* vol. 12, no. 4, p. 149, 2004.

[48] C. Gussenhoven, “Intonation and interpretation: phonetics and phonology,” in *Speech Prosody 2002*, Aix-en-Provence, France, 2002.

[49] R. Hincks and J. Edlund, “Stimulating Increased Pitch Variation in Oral Presentations with Transient Visual Feedback,” *Language Learning and Technology.,* 2008.

[50] M. E. Hoque, “Computers to Help with Conversations: Affective Framework to Enhance Human Nonverbal Skills,” *PhD Dissertation, Massachusetts Institute of Technology,* 2013.

[51] J. Kim, “Automatic Detection of Sentence Boundaries,” *PhD Dissertation, Dept. Elect. Eng., University of Washington.,* 2004.

[52] R. Ranganath, D. Jurafsky and D. A. & McFarland, “Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates,” *Computer Speech & Language.,* vol. 27, no. 1, pp. 89-115, 2013.

[53] K. S. Rao and S. G. Koolagudi, Robust Emotion Recognition using Spectral and Prosodic Features, Springer Science & Business Media, 2013.

[54] A. Ritchart and A. Arvaniti, “The form and use of uptalk in Southern Californian English,” in *Social and Linguistic Speech Prosody: Proceedings of the 7th international conference on Speech Prosody*, Dublin, 2014.

[55] B. Schuller and A. Batliner, Computational paralinguistics: emotion, affect and personality in speech and language processing, John Wiley & Sons, 2013.

[56] V. Sethu, E. Ambikairajah and J. Epps, “Pitch contour parameterisation based on linear stylisation for emotion recognition,” *INTERSPEECH,* pp. 2011-2014, 2009.

[57] R. Signorello, F. D'errico, I. Poggi and e. al, “How Charisma Is Perceived from Speech: A Multidimensional Approach,” in *Privacy, Security, Risk and Trust (PASSAT), 2012 International*

Conference on and 2012 International Conference on Social Computing (SocialCom), Amsterdam, 2012.

[58] P. A. Taylor, "Automatic recognition of intonation from F0 contours using the rise/fall/connection model," *ATR Interpreting Telecommunications Laboratories,* 1993.

[59] A. Veiga, D. Celorico, J. Proença and e. al, "Prosodic and phonetic features for speaking styles classification and detection," in *Advances in Speech and Language Technologies for Iberian Languages*, Springer Berlin Heidelberg., 2012, pp. 89-98.

[60] D. Wang and S. S. Narayanan, "Robust speech rate estimation for spontaneous speech," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, no. 8, pp. 2190-2201, 2007.

[61] C. E. Williams and K. N. Stevens, "Emotions and speech: Some acoustical correlates," *The Journal of the Acoustical Society of America,* vol. 52, no. 4B, pp. 1238-1250, 1972.

[62] Y. Zhang and J. Glass, "Speech rhythm guided syllable nuclei detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2009.

[63] M. E. Beckman and J. Hirschberg, "The ToBI Annotation Conventions," [Online]. Available: http://www.ling.ohio-state.edu/~tobi/ame_tobi/annotation_conventions.html. [Accessed 3 June 2015].

Appendix A: Speech Analysis Values

Name	Group	Length (s)	F0mean	F0std	Pdur_mean	Pstd	P%	Udur_mean	Ustd	U%	HRT%	HRTnumHRT	FI%	FInumFI
Student 1	PSY2031	272	185	46	0.592	0.189	0.20	2.304	2.219	0.80	0.15	11.00	0.08	6
Student 2	PSY2031	189	185	36	0.548	0.128	0.18	2.443	5.655	0.82	0.02	1.00	0.02	1
Student 3	PSY2031	310	165	47	0.531	0.367	0.15	3.026	4.137	0.85	0.14	11.00	0.18	14
Student 4	PSY2031	234	157	42	0.594	0.175	0.13	3.879	11.598	0.87	0.07	3.00	0.12	5
Student 5	PSY2031	232	223	42	0.930	0.809	0.23	2.989	5.049	0.77	0.18	9.00	0.02	1
Student 6	PSY2031	277	162	46	0.853	1.200	0.25	2.507	4.813	0.75	0.12	6.00	0.12	6
Student 7	PSY2031	311	167	44	0.775	0.368	0.22	2.715	3.804	0.78	0.42	30.00	0.17	12
GROUP AVERAGE		261	178	43	0.689	0.462	0.20	2.837	5.325	0.80	0.16	10.14	0.10	6.43
AimeeMullin	TED	300	209	37	0.650	0.148	0.28	1.659	1.688	0.72	0.01	1.00	0.05	4
AngelaDuckworth	TED	300	262	73	0.724	0.142	0.25	2.109	1.231	0.75	0.07	7.00	0.63	59
EmilyBalcetis	TED	300	249	67	0.659	0.173	0.21	2.470	2.405	0.79	0.11	9.00	0.34	28
HelenFisher	TED	300	215	46	0.500	0.095	0.12	3.791	8.169	0.88	0.07	4.00	0.16	10
JanineShepherd	TED	300	207	41	0.933	0.918	0.23	3.077	8.915	0.77	0.03	2.00	0.03	2
KellyMcGonigal	TED	300	223	56	0.993	1.346	0.29	2.381	2.708	0.71	0.01	1.00	0.40	29
PamelaMeyer	TED	300	192	55	0.757	0.201	0.28	1.952	1.337	0.72	0.13	12.00	0.18	16
GROUP AVERAGE		300	222	53	0.745	0.432	0.24	2.491	3.779	0.76	0.06	5.14	0.26	21.14
Miranda	WEDDING	300	233	43	1.057		0.26	2.898	2.768	0.74	0.00	0.00	0.14	10
Nate	WEDDING	300	118	18	1.203	0.891	0.39	1.863	1.077	0.61	0.00	0.00	0.03	2
Oli	WEDDING	300	145	29	1.606	1.499	0.42	2.182	2.357	0.58	0.02	1.00	0.16	10
Riva	WEDDING	300	157	30	0.858	0.508	0.30	2.005	1.085	0.70	0.00	0.00	0.29	25
Ruby	WEDDING	300	126	19	0.667	0.452	0.20	2.666	2.359	0.80	0.00	0.00	0.06	5
Tara	WEDDING	300	216	40	0.490	0.158	0.16	2.474	2.058	0.84	0.00	0.00	0.21	19
GROUP AVERAGE		300	166	30	0.980	0.702	0.29	2.348	1.951	0.71	0.00	0.17	0.15	11.83
PSY2031 AVERAGE		261	**178**	**43**	**0.689**	0.462	**0.20**	2.837	**5.325**	**0.80**	**0.16**	10.14	**0.10**	6.43
TED AVERAGE		300	**222**	**53**	**0.745**	0.432	**0.24**	2.491	**3.779**	**0.76**	**0.06**	5.14	**0.26**	21.14
WEDDING AVERAGE		300	**166**	**30**	**0.980**	0.702	**0.29**	2.348	**1.951**	**0.71**	**0.00**	0.17	**0.15**	11.83

Length: duration of audio clip analysed, Pdur_mean: mean pause duration, Pstd: Pause Standard Deviation, P%: percentage of recording which is pauses
Udur_mean: mean utterance duration, U%: percentage of recording which is speech, HRT%: percentage of utterances ending in HRT, FI%: percentage of utterances ending in FI

Appendix B: Requirements

ECE4094: Requirements Analysis

Ruben Bloom – 21507252
25 April 2014

REVISION CONTROL

Version	Date	Details
1.0	28-Apr-14	First Draft
1.1	29-Apr-14	Modified based on comments from Wai Ho Li

OBJECTIVES

The objective of this project is to design a software application which informs a user about the paralinguistic features of their speech such as pitch, speech rate, disfluencies, and number of stressed syllables. Paralinguistic features capture a user's speaking style. S*peaking style* refers to the manner of speaking which communicates the personality traits, affective states, and attitudes of the speaker.

The application will allow a user to compare their speech's paralinguistics features, and by extension their speaking style, to those of other speakers and of themselves at previous times.

It is anticipated that this tool will have the following applications:

- By providing quantifiable measures of paralinguistic speaking style, the application will assist a speaker to modify their speaking style towards a desired style.
- The application will assist voice and speech coaches in assessing and training their clients.
- The application may be used in research.
    - Analyses of speech samples from speakers determined to be charismatic and uncharismatic may allow for quantification of the speech characteristics which contribute to charisma.
    - Comparison of speech samples between healthy and non-healthy subjects can be used to characterise and diagnose conditions affecting speech such as amyotrophic lateral sclerosis.

BACKGROUND

Consciously and unconsciously speakers embed significant information about their affective state, personality traits, and intentions in the paralinguistic features of their speech such as pitch and speech rate [1].
This information strongly affects how a speaker is perceived and received by listeners.

Paralinguistics features of speech which influence how a speaker is perceived have acoustic correlates which can be extracted from the speech waveform. To date, several studies have explored the acoustic correlates of traits inferred from speech such as convincingness and power [25], liveliness [26], insecurity, hesitance, trustworthiness, humility, [27], passion, anger, and intensity [28].

While focused on emotion detection, the field of affective computing has generated a sizeable corpus of research on the analysis of paralinguistic acoustic features from speech [29] [2]. Computational paralinguistics is now emerging as a new field within speech and language processing [30].

Currently there exist speech analysis toolkits designed for researchers such as Speech Filing System [10] and Praat [11], however these toolkits do not implement analysis for high level paralinguistic features such as speech rate and disfluencies and are not designed for use by a non-expert user.

This project will use the speech feature analysis research from affective computing and computational paralinguistics to create a useful tool for non-expert users to understand and modify their speaking style.

TYPES OF REQUIREMENTS

There are several distinctions made between the different types of requirements:

1. **Requirements.** Standard requirements that are necessary to fulfil in order for the product to meet customer expectations. These are listed in the form "R.xxx"

2. **Design Guides.** Customer-supplied suggestions relating to any aspect of the design of the product. These are listed in the form "DG.xxx"

3. **Optional.** Standard requirements that the customer has indicated are to be fulfilled if possible, with the expectation that they might not be possible or feasible to fulfil. These are listed in the form "OR.xxx".

4. **Caveats.** Specifies limitations imposed on the system put in place in order to meet the requirements. They are listed in the form “C.xxx”.

REQUIREMENTS

The software application, hereafter referred to as the 'the system', is envisioned to consist of a graphical user interface through which a user uploads a speech file or records a new sample; software which processes the speech sample; and a graphical interface displaying the computed speech features.

[R.001] The system will be accompanied by a literature review summarising research and techniques relevant to the system to be developed.

[R.002] The system must analyse a speech sample from a user and extract at least four paralinguistic speech features such as pitch, speech rate, and number of pauses. The combination of these results constitute the speaking style in the given sample.

[R.011] The system will work with any recording of a quality equivalent to at least 16bits per sample, 16,000Hz sample rate and of reasonable SNR (to be determined empirically).

[R.012] The speech sample may be a recording of a user speaking a predetermined passage.

[OR.012] Optionally, if a more advanced system is feasible than the speech sample used may be any short speech sample from a user.

[R.003] The selected speech features must have evidence of being involved in human perception of speaking style. Evidence can be sourced from research in affective computing, relevant social sciences, voice coaching practices, speech pathology practices, or empirical research.

[R.004] Each speech feature measurement must be checked for validity against an external measure, e.g. manual transcription of speech rate or pauses, comparing pitch analysis with alternative algorithms.

[R.005] The system must be able to reliably indicate a difference in the speaking styles of two speech samples if humans rate the samples as having significantly different speaking styles.

[OR.005] If there are cases in which the features measured in the system fail to reflect differences in speaking style which humans consistently detect, an investigation will be conducted.

[R.006] The system will be accompanied by an explanation of the quantities measured which is meaningful to the user. For instance, an explanation of what pitch is and it is relevance to perceived speaking style.

[R.007] The results of the analysis must be displayed graphically.

[R.008] The user interface of the system must allow the user to compare at minimum two different speaker profiles.

[R.009] The analysis will take no longer than 120s for a 30s speech sample on a PC with 1GHz or faster processor, 1GB RAM, and 16GB available hard disk space.

[OR.009] The processing may be moved to a server to meet real time requirements.

[R.010] The end system will be an application usable on a widely available device type and operating system.

[DG.010] The end system may take the form of a stand-alone Windows application, compatible with Windows 7 32-bit OR a web-application designed to be usable by a browser with either HTML5, Adobe Flash Support, or WebRTC depending on implementation. If a web-application, the system will be guaranteed to work successfully on at least one web browser. OR, the system may be implemented on the Android smartphone operating system with either local or server processing.

[C.001] The system will be designed and tested with native English speakers. Full functionality and reliability are not guaranteed with speech samples from other languages or from non-native English speakers.

References

[1] A. Mokharti and N. Campbell, "Speaking Style Variation and Speaker Personality," *Proc. of Speech Prosody,* 2008.

[2] D. Braga and M. A. Marques, "The pragmatics of prosodic features in the political debate," in *Speech Prosody 2004, International Conference*, 2004.

[3] R. Hincks, "Computer support for learners of spoken English," *Diss. Speech and Music Communication, KTH,* 2005.

[4] E. Strangert, "What makes a good speaker? Subjective ratings and acoustic measurements," in *Proceedings from Fonetik 2007: speech, music and hearing, quarterly progress and status report*, 2007.

[5] J. B. Hirschberg and A. Rosenberg, "Acoustic / Prosodic and Lexical Correlates of Charismatic Speech," in *Proceedings of Eurospeech'05*, Lisbon, Portugal, 2005.

[6] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International Journal of Speech Technology,* pp. 99-117, 2012.

[7] Z. Zeng, M. Pantic, G. I. Roisman and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Tansactions on Pattern Analysis and Machine Intelligence,* vol. 31, no. 1, pp. 39-58, 2009.

[8] B. Schullera, S. Steidlb, A. Batliner, F. Burkhardtd, L. Devillersa, C. Müllere and S. Narayananf, "Paralinguistics in Speech and Language--State-of-the-Art and the Challenge," *Computer, Speech, and Language,* pp. 4-39, 2013.

[9] M. Huckvale, "Speech Filing System. Tools for speech research," 2014. [Online]. Available: http://www.phon.ucl.ac.uk/resource/sfs/.

[10] P. Boersma and D. Weenink, "Praat: doing phonetics by computer.," 2014. [Online]. Available: http://www.praat.org/.