

**UC Berkeley**

**UC Berkeley Electronic Theses and Dissertations**

**Title**

Automated Detection of Social Signals Using Acoustic and Lexical Features of Extemporaneous Speech in Naturalistic Environments

**Permalink**

<https://escholarship.org/uc/item/7h75g8xd>

**Author**

Corrigan, Seth

**Publication Date**

2022

Peer reviewed|Thesis/dissertation

Automated Detection of Social Signals Using Acoustic and Lexical Features of Extemporaneous  
Speech in Naturalistic Environments

By

Seth Corrigan

A dissertation submitted in partial satisfaction of the

Requirements for the degree of

Doctor of Philosophy

in

Education

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Mark Wilson, Chair

Professor Karen Draney

Professor Zach Pardos

Summer 2022

**Automated Detection of Social Signals Using Acoustic and Lexical Features of  
Extemporaneous Speech in Naturalistic Environments**

Copyright 2022

by

Seth Corrigan

## Abstract

### Automated Detection of Social Signals Using Acoustic and Lexical Features of Extemporaneous Speech in Naturalistic Environments

by

Seth Corrigan

Doctor of Philosophy in Education

University of California, Berkeley

Professor Mark Wilson, Chair

The goal of this dissertation is to study and develop approaches to automating detection of social signals from speech using extemporaneous talk gathered in naturalistic settings. All three chapters focus on detection of what is referred to as competence-focused and likability-focused speech, which are two examples of social stances that humans advance in social interactions and that may be detected by machines.

The first chapter describes the development and performance of an approach to detecting competence-focused and likability-focused speech among expert speakers, namely, professional actors and voice-over experts. I demonstrate that speakers' attempts to advance such social stances can be detected with a level of accuracy that approximates an existing benchmark. The second chapter follows a similar design and approach but uses instead a corpus of audio recordings collected from non-expert speakers—participants who do not have training or experience as actors.

The first and second chapters describe models that were developed and tested to use the acoustic features of recorded speech to infer whether the speaker was responding to a social situation and directive that prompted competence-focused speech or likability-focused speech. In those cases, the classification problem required an inference about the stimulus that prompted the speaker. There is also merit in inferring how a human interlocutor would perceive a given sample of speech, an example of a general type of problem that has been termed inferential detection. Inferential detectors utilize machine learning and measurement processes to infer human judgements of objects, agents, processes, or environments even in the absence of a human observer. The third chapter presents a general process for developing inferential detectors of social stances. In addition, I develop and describe inferential detectors for competence-focused and likability-focused speech that utilize multiple sources of information about the speaker, in this case, acoustic features of speech as well as its lexical content.





For Coach, the rest of my wonderful family, and my teachers.

## Table of Contents

CHAPTER 1	
CHAPTER ONE .....	1
Detection of Social Signals Among Expert Speakers.....	1
Purpose.....	1
Intended Contributions.....	1
Background.....	2
Social Stance: Likability and Competence .....	2
Acoustic Features of Speech.....	2
Past Efforts at Detection of Social Stances.....	3
Materials and Methods.....	5
Study Design.....	5
Participants.....	5
Speech Recording .....	6
Speech Pre-Processing.....	7
Data structure and Data Processing .....	7
Outcome Definition .....	7
Analysis.....	7
Data Preparation.....	8
Data Partitions.....	8
Feature Importance .....	8
Modeling Approach .....	9
The L1 Logistic Regression Classifier.....	9
Optimization of the L1 Logistic Regression Classifier.....	10
The Support Vector Machine (SVM) .....	10
Optimization of the Support Vector Machine.....	12
Model Evaluation.....	12
Unweighted Average Recall .....	12
Area Under the Curve .....	13
Sensitivity and Specificity .....	14
Results.....	14
Variable Importance.....	14
Model Performance.....	16
Discussion.....	17

Characteristics of Speech Sounds .....	17
Model Results .....	17
Conclusion .....	18
CHAPTER TWO .....	19
Detection of Social Signals Among Non-Expert Speakers.....	19
Purpose.....	20
Intended Contributions.....	20
Background.....	20
Social Signals and Prosodic Features of Speech.....	21
Solving Coordination Problems in Small Windows of Time .....	21
Solving Coordination Problems Across Larger Windows of Time .....	22
Multimodal Communication.....	22
Materials and Methods.....	24
Study Design.....	24
Participants.....	25
Speech Recording .....	26
Speech Pre-Processing .....	27
Feature Extraction via openSMILE .....	27
Outcome Definition .....	28
Analysis.....	29
Data Preparation.....	29
Data Partition .....	29
Feature Importance .....	29
Modeling Approach .....	30
The L1 Logistic Regression Classifier.....	30
Optimization of the L1 Logistic Regression Classifier.....	31
The Support Vector Machine (SVM) .....	31
Optimization of the Support Vector Machine.....	32
Model Evaluation.....	33
Unweighted Average Recall .....	33
Area Under the Curve .....	34
Sensitivity and Specificity .....	35
Results.....	35
Variable Importance.....	35

Model Evaluation.....	38
L1-Logistic Regression Performance.....	38
Support Vector Classifier Performance .....	39
Support Vector Machine Performance.....	39
Discussion.....	39
Patterns in Speaker Behavior .....	39
Energy.....	40
Audio Spectrum .....	40
Patterns in Model Performance.....	41
Strengths and Limitations .....	41
Conclusion .....	42
CHAPTER THREE .....	44
Use of Acoustic and Lexical Features of Speech to Detect Social Signals .....	44
Purpose.....	44
Intended Contributions.....	45
Background.....	45
The Source-Filter Theory and Formant Frequencies .....	47
Double Articulation .....	49
Use of Acoustic and Lexical Features in Speech to Convey Social Signals .....	49
Inferential Detectors.....	50
A General Development Process for Inferential Detectors.....	50
Inferential detectors for Competence-Focused and Likability-Focused Speech .....	53
Materials and Methods.....	53
Study Design.....	53
Participants.....	53
Speech Recordings.....	54
Speech Pre-Processing.....	54
Audio-Clip Selection .....	54
Audio Clip Review and Rating.....	54
Data Structure and Data Processing.....	55
File format for the Classification Task .....	56
Outcome Definition .....	57
Analysis.....	57
Data Preparation.....	57

Sampling, Rating, and Labelling Procedures.....	58
Data Partitions.....	59
Modeling Approach .....	60
The Faceted Rating Scale Model (FRSM).....	60
Notes on the Use and Interpretations of the FRSM in the Current Study.....	61
Feature Importance .....	62
Mutual Information.....	63
Supervised Machine Learned Models.....	63
Performance Metrics.....	66
Unweighted Average Recall .....	67
Area Under the Curve .....	67
Results.....	69
FRSM Results .....	69
Variable Importance Results.....	73
Classification Model Results .....	81
Discussion.....	82
Speaker Behaviors .....	82
Use of the Faceted Rating Scale Model for Inferential Detectors .....	83
Leveraging Multiple Sources of Information .....	83
Future Work.....	84
Conclusion .....	84
Appendix A.1: R Scripts for Chapter 1.....	86
Appendix A.2: Sample Recording Tasks.....	94
Appendix A.3: eGeMAPS Acoustic Parameters .....	98
Appendix B: R-Scripts for Chapter 2.....	101
Appendix C.1: R Scripts for Chapter 3.....	114
Appendix C.2: Estimated Facets for Competence-Focused Speech.....	114
Appendix C.3: Estimated Facets for Likability-Focused Speech.....	114
References.....	115

## CHAPTER ONE

### Detection of Social Signals Among Expert Speakers

Speakers in conversation engage in complex behaviors that convey socially relevant signals and influence the social situations in which they participate. This requires successful management of the impressions and inferences of others and involves enacting one or more social stances (Ochs, 1993) by coordinating the propositional content of utterances (*what one says*) as well as their acoustic features (*how one says it*). In this regard, enacting a social stance can be understood to involve a type of communicative competence (Wieman, 1977; Canale & Swain, 1981). Variations in communicative competence can be identified with varying levels of success on the part of speakers in managing their communicative behavior, influencing the impressions of others, exhibiting empathy, inviting appropriate levels of affiliation, and showing support (Wieman, 1977).

Two stances that have relevance for everyday social interactions and individuals' life chances involve emphasizing one's likability (i.e., one's readiness for affiliation with others) and competence (i.e., one's ability to accomplish joint action or harm another). At least two general benefits accrue from description and detection of the communicative behaviors people use to portray stances such as likability and competence. First, a detailed description of communicative behavior is expected to facilitate automatic detection of social stances—an outcome that stands to contribute to improving human-computer interactions (HCI) and even human-human interactions. Second, data-driven descriptions of the communicative behavior associated with varying levels of communicative competence in general, and stance-taking in particular, is expected to benefit both future work in automated detection of social stances and efforts to support individual development of the associated sociolinguistic skills.

#### Purpose

The purpose of the current study is two-fold: first, the study aims to investigate the feasibility of automating detection of likability-focused and competence-focused speech when speech production is unscripted and occurring in uncontrolled environments; second, the study aims to identify and describe acoustic features that are most productive in distinguishing likability-focused speech from competence-focused speech. Importantly, the work is carried out using supervised machine learning methods that contribute to automated detection of the two target stances, likability and competence, using acoustic features of speech.

Voice recordings of extemporaneous monologues were gathered from a group of actors. The sample was well balanced with regard to reported sex at birth and all of whom are native English speakers living within the continental United States. A series of machine-learned models were developed using L1 logistic regression and a support vector machine. The models' performances are evaluated for their accuracy in differentiating speech associated with the likability-focused and the competence-focused recording tasks. Acoustic features of speech that are identified as contributing unique information to the classification of likability-focused and competence-focused speech among expert and non-expert speakers are used to describe and compare variation in communicative behavior.

#### Intended Contributions

While there exists a substantial body of literature on automated detection of likability-focused speech (e.g., Cummins et al., 2012; Gonzales et al., 2013; Ranganath et al., 2013; Schuller et al., 2015), less work exists on detecting *competence-focused* speech and identifying the features with which it is associated. This work seeks to contribute to an understanding of the

features of both likability-focused and competence-focused speech and automation of their detection.

### **Background**

Social stances are complex sets of communicative behaviors used by speakers to provide signals about who they are (identity), their commitments to what they are saying (epistemic), and their willingness to affiliate with others (social) (Ochs, 1992, 1993; Kiesling, 2009). During social interactions, speakers control their communicative behavior to intentionally portray one or more social stances (Ochs, 1992) as a means to manage their interlocutors' inferences and impressions (Goffman, 1959; Garfinkel, 1967). In turn, interlocutors utilize the resulting social signals to make inferences about the speaker, ensure the interaction transpires smoothly, and to make repairs when misunderstandings arise (Schegloff et al., 1977).

### **Social Stance: Likability and Competence**

Research on social stance and its detection intersects with work in social psychology and social perception that aims to describe how individuals portray and convince others of their likability and competence—two examples of social stances that are critical to successful social and professional interaction.

Inferences about the likability and competence of others have been hypothesized to be basic and universal (Fiske et al., 2007; Cuddy et al., 2008; Fiske, 2018). From the standpoint of social psychology, the need to quickly infer another's intentions (harmful vs. cooperative) and their capacity to follow through on those intentions (capable vs. incapable) is at the root of many social stereotypes (Fiske et al., 2007), social norms, and aspects of human development (MacDonald, 1992).

An ability to successfully control features of one's speech to portray likability and competence has implications for a range of life outcomes. These include contexts in which individuals' likability and competence are under intense scrutiny in professional settings such as job interviews (Gilmore et al., 1999) and salary discussions (Curhan & Pentland, 2007). Similarly, there is a growing body of evidence that properties of speech are associated with one's choice of intimate partners (Oguchi & Kikuchi, 1997; Collins, 2000; Feinberg et al., 2005), who one does business with (Burkhardt et al., 2011; Parhankangas & Ehrlich, 2014), who one votes for (Tigue et al., 2012; Klofstad et al., 2012), and who one trusts (Levitan et al., 2018).

### **Acoustic Features of Speech**

While production of social stances that portray likability and competence utilize the propositional content of speech, successful portrayal also relies on acoustic features. Acoustic features of speech can be understood as sets of parameters contributing to a complex dynamic speech system (De Bot et al., 2007; Verspoor, 2013, 2017; MacIntyre and Ayers-Glassey, 2020). They relate to the physical properties of the sound waves associated with a sample of speech. With many of those properties under varying levels of control by the typical speaker, they are used with varying degrees of success to influence impressions of one's likability and competence.

In the context of social encounters, speakers exhibiting communicative competence are generally expected to adjust their communicative behavior in ways that emphasize or deemphasize their likability and/or competence as required by the given communication goal and setting (Eckert & Rickford, 2001; Regan, 2010; Geeslin et al., 2018). Speakers who exhibit greater control over relevant acoustic properties of their speech may be more successful at managing the impressions of others. Control in this regard may point to both the number and types of behavioral patterns speakers invoke.



### Past Efforts at Detection of Social Stances

Early work on social stances (Leary, 1957; Ochs, 1992) has more recently been leveraged for development of automated detectors capable of classifying speech into one or more categories of stances. Many of these efforts aim to improve social services such as policing and psychiatry. Others aim to improve human-machine interaction and to provide tech-based tools for teaching and learning. Efforts to automate detection of individuals' stances also result in identification of the features of speech likely to aid human listeners' detection of the targeted stances. While there are many examples of such work that focus on likability or an associated construct, fewer instances are available that focus on detection of competence or related constructs.

Efforts to leverage acoustic and lexical features of speech for automating detection of speaker likability or competence are often approached as two-class problems in which the target stance is either present or absent, and they tend to utilize nonlinear supervised classification approaches such as support vector machines (*e.g.*, Cummins et al., 2012; Gonazales & Anguera, 2013; Schuller et al., 2015). Past efforts vary in the type of stimuli they use to prompt the recorded speech and the extemporaneity of the speech. It is typical to use short, scripted speech samples that are 1-2 sentences in length and collected from a relatively small number of speakers (*e.g.*, Schuller et al., 2012). They also vary in the type and number of speaker-participants they engage – whether the speakers are non-experts or actors, recording speech on their own, in pairs or in larger groups.

When the length of the resulting recordings is sufficiently long, researchers also exhibit a range of choices in the window size at which the recordings are analyzed, or how they will summarize across lengthier windows of time. In cases in which multiple speakers are engaged, researchers are also tasked with identifying turns-of-talk so that speech from each individual can be treated separately if desired. Examples of past work to detect one or more stances are described here in more detail with attention given to the constructs and modeling approaches used.

Burkhardt et al. (2011) automate detection of speaker characteristics such as sex and age, as well as likability. They use scripted recordings from 100 nonexpert speaker participants, each recording eighteen utterances. Participants rated recordings for likability using a seven-point scale and a majority vote was utilized as the label for each audio recording. Using an ensemble of Random Forests, the group achieved an unweighted average accuracy for the two-class classification problem (*likable v not likable*) of 67.6%.

Schuller et al. (2012), via the INTERSPEECH 2012 Speaker Trait Challenge, later spurred numerous efforts to detect likability-focused speech. In that INTERSPEECH challenge, they shared a dataset, acoustic feature set, and described baseline model performance to allow others to gauge their own success with the classification task. Speech data for the competition came from Burkhardt et al. (2011). The feature set included acoustic features only making up the Computational Paralinguistics Challenge (ComParE) feature set<sup>1</sup> with over 6,000 low level and

---

<sup>1</sup> The ComParE feature set is a large set of acoustic features, or properties of speech that were identified for the 2011 INTERSPEECH Challenge. The feature set is described in detail in Eyben et al. (2010); the authors also introduce the available Python scripts that can be used to identify and extract low-level features of speech recordings and functional features that result from application of one or more mathematical functions to the low-level features. The available Python scripts are a part of the open Speech and Music Interpretation by Large Space Extraction (openSMILE) library which makes available scripts that permit identification and extraction of acoustic features of speech and music. Low-level features are those acoustic properties that can be measured directly, such as loudness. The functional features are those properties of speech that are derived, such as rate of speech, which is a count of

functional features. The baseline model was a linear logistic regression, which they compared with a support vector machine (Boser et al., 1992) and use of a Random Forest (Breiman, 2001). The random forest model performance was best with an unweighted average accuracy of 59% and an AUC of 64.7.

Ranganath et al. (2013) utilized support vector machines in the context of speed dating recordings to detect flirtatiousness and friendliness using acoustic and lexical features of speech. Speech data for the competition came from a series of speed dates organized for the study with both participants in each speed date using a separate microphone. The feature set included low level acoustic features of the daters' speech, lexical features with additional features that captured patterns in the speech arising across both daters. The team used L1 logistic regression as their baseline model, which they compared with a support vector machine (SVM). The SVM outperformed the L1 logistic regression with an accuracy of 58.0% for detecting friendliness.

As mentioned previously, limited work has been pursued to automate detection of competence-focused speech, though there exist efforts to detect potentially related stances. Formolo and Bosse (2017, 2018), for example, aim to improve human-computer interactions by automating detection of dominance among seven other stances as portrayed through speech. The team recorded 24 scripted sentences from twenty actors, ten male and ten female, with all 24 sentences recorded multiple times by each actor, once for each of the eight interpersonal stances: *leading, helping, cooperating, dependent, withdrawn, aggressive, defiant, and competitive*. They used the OpenSMILE Feature Set from the 2011 INTERSPEECH Paralinguistic Challenge (Eyben et al., 2013) and employed an SVM for each of the eight stances. The team evaluated the approach and found the average unweighted average accuracy in classifying the eight stances was 19.3%, only slightly better than chance (12.5%).

Additional work to detect social stances that may relate to portrayal of competence is presented by Pon-Barry and Shieber (2011) who developed a detector for speakers' self-perceived level of uncertainty. They utilized primarily prosodic features of speech to detect speaker uncertainty. The researchers engaged twenty actors to read thirty scripted sentences in a cloze format, requiring the actors to select a word that best completed the sentence. The cloze tasks were constructed in a manner that was designed to elicit varying levels of uncertainty regarding how to best complete each sentence. After each sentence, actors rated their own level of uncertainty on a five-point scale ranging from *very uncertain* to *very certain*. The researchers trained a single decision tree to classify the recordings into each of the scale categories for the actors' self-perceived ratings and achieved an accuracy of 63.3%. Using a set of decision trees, one for each set of stimuli (those intended to support certainty and those intended to support uncertainty), the team achieved an overall accuracy of 75.3%.

There exist several other examples of such detectors. As suggested here, most utilize one or more ordinal rating scales that speaker-participants or reviewer-participants use to associate a quantitative value with their perception of the audio recordings—primarily, the extent to which they feel the given clip exhibits the targeted stance. These values then serve as ground truth for subsequent work to develop one or more classifiers. In most cases, this has the potential to introduce an unknown, unspecified source of variance that has its origin in individual differences that are not accounted for.

In one view, the human raters in these studies may be understood to be performing the role of transducers, transducing the acoustic signals they perceive in the recordings into

---

words spoken per unit time. The OpenSMILE Python library is freely available (<https://audeering.github.io/opensmile-Python/>).

information in the form of one or more rating scale values. When several transducers generate different rating scale values that vary in unsystematic ways for the same recording or set of acoustic properties, the resulting ratings cannot be compared. While work by Pon-Berry and Shieber (2011) provides additional design elements that may alleviate the problem, access to criterion-referenced scales for the various stances of interest may present further benefits. It is expected that when raters have access to qualitative descriptions of the properties of speech at varying levels of likability, for example, that it would put raters' scores on a more objective footing and allow for more consistent use of the rating scales.

While several studies have focused on identifying the features of speech associated with a range of social stances, it is not clear that there is widespread agreement on what those features are and for whom. Additional work to identify features of speech associated with social stances is warranted. Further, while most efforts utilize scripted utterances to generate a bank of recordings, there may be benefits derived in the form of greater generalizability from using extemporaneous speech. Research in the field of sociolinguistics points to the need to gather extemporaneous speech that arises in naturalistic environments, as stance is viewed as something that is constructed through interactions with others and even objects in the environment. Identifying methods for compiling, and working with corpora of extemporaneous speech, may be of help in improving out-of-sample performance of detectors for speech data. Lastly, as mentioned at the outset of the proposal, although there are many examples of research efforts to automate likability of speech using acoustic features, there are fewer examples of such efforts for competence and related constructs. In what follows an approach is described for detecting social stance from extemporaneous speech in naturalistic settings. Detection of likability-focused and competence-focused speech are both investigated. Importance values are estimated for each feature in order to identify those acoustic features most effective in distinguishing competence-focused and likability focused speech.

### **Materials and Methods**

As a means to promote transparency and openness, access to source code for analyses is provided in Appendix A.1.

#### **Study Design**

This study describes development of an automated detector for likability-focused and competence-focused speech from expert speakers responding extemporaneously to prompts in naturalistic environments. The aim is two-fold: first, the study aims to investigate the feasibility of automating detection of likability-focused and competence-focused speech when speech production is unscripted and occurring in uncontrolled environments; second, the study aims to identify and describe acoustic features that are most productive in distinguishing likability-focused speech from competence-focused speech. The study utilizes a cross-sectional design with data collected from crowd-sourced participants.

#### **Participants**

Audio recordings from expert speaker-participants with training or professional acting experience were gathered and used for the study. A total of 101 adults (48 (47.5%) female and 53 (52.5%) male), 20-65 years of age ( $M = 40.6$ ,  $SD = 12.6$ ), participated as expert speakers in this study. All speaker participants were native speakers of American English and were currently residing in the United States. A summary of participants' reported sex at birth and racial/ethnic backgrounds is provided in Table 1.1.

**Table 1.1**  
*Demographics Reported by Participating Expert Speakers*

	American Indian		Asian		Black		Pacific Islander		White		Multi-ethnic		Decline		Other		Total	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
All	0	0.00	4	3.7	13	12.9	0	0.	77	76.2	5	5.0	2	2.0	0	0.0	101	100.0
<b>Female</b>	<b>0</b>	<b>0.00</b>	<b>4</b>	<b>4.0</b>	<b>9</b>	<b>8.9</b>	<b>0</b>	<b>0.0</b>	<b>33</b>	<b>32.7</b>	<b>2</b>	<b>2.0</b>	<b>0</b>	<b>0.0</b>	<b>0</b>	<b>0.0</b>	<b>48</b>	<b>47.5</b>
Hispanic	0	0.00	0	0.0	9	8.9	0	0.00	6	5.9	0	0.0	0	0.0	0	0.0	15	14.9
Non-Hispanic	0	0.00	4	4.0	7	6.9	0	0.0	27	26.7	2	2.0	0	0.0	0	0.0	40	39.6
<b>Male</b>	<b>0</b>	<b>0.00</b>	<b>0</b>	<b>0.00</b>	<b>4</b>	<b>4.0</b>	<b>0</b>	<b>0.0</b>	<b>44</b>	<b>43.6</b>	<b>3</b>	<b>3.0</b>	<b>2</b>	<b>2.0</b>	<b>0</b>	<b>0.0</b>	<b>53</b>	<b>52.5</b>
Hispanic	0	0.00	0	0.0	1	1.0	0	0.0	1	1.0	0	0.0	2	2.0	0	0.0	4	4.0
Non-Hispanic	0	0.00	0	0.00	3	3.0	0	0.0	43	42.6	3	3.0	0	0.0	0	0.0	49	48.5

Expert speakers were recruited using the online markets specializing in gig work for professional actors and voice-over experts<sup>2</sup>. Once candidates responded to communications on those sites describing the study, candidates' biographies were reviewed for experience and formal training in voice-over or acting. Candidates' biographies were also reviewed to ensure they were native speakers of American English currently residing in the United States. All speaker participants received remuneration for taking part in the study.

### Speech Recording

Participants were given access to the Online Recording System (ORS), a cloud-based authoring tool for graphic novel experiences which integrates audio recording functionality. The ORS requests access to the microphone on the user's computer and leads the user through a series of checks to ensure the microphone is working correctly and that the noise-level in the recording environment is acceptable. Users are then introduced to the story line of a graphic novel titled Advice Hour, and participants are invited to assume the role of a podcast host responding to callers' questions about how to handle specific communication dilemmas in their personal and professional lives. Each of the scenarios ends with a prompt for the speaker-participant to record what they would want to say in the given scenario in the manner in which they would say it. The recording prompts used for this research explicitly request either competence-focused or likability-focused speech. A sample of the recording prompts is included in Appendix A.2. As a part of the ORS functionality, after each response, participants are given the chance to review their recording and either accept or revise it. The ORS stores participant-accepted recordings in a secure cloud-based file.

<sup>2</sup> Two online markets were used in the current study: Fiverr (Fiverr; <http://fiverr.com>), and Upwork (<http://Upwork.com>).

## Speech Pre-Processing

Members of the research team reviewed speaker-participants' recordings for evidence of on-task performances. Recordings for each task were then segmented into five-second clips<sup>3</sup> which were associated with a unique identifier allowing the task, task type (likability-focused task vs. competence-focused task), and window rank of the recording (the cardinal value denoting the position of the clip in the full recording, such as *first 5s window*, *second 5s window*, etc.) to be indexed.

## Data structure and Data Processing

The Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) parameter set was extracted from participants' five-second clips of speech for each task using the OpenSMILE feature library (Eyben et al., 2013)<sup>4</sup>. The eGeMAPS parameter set provides arithmetic means and coefficients of variation (standard deviation (S.D. normalized by the arithmetic mean) for each acoustic feature. All 88 parameters in the set were computed for each five-second clip for each task and speaker combination. The resulting set of parameter values were used for the model building efforts.

A full overview of the parameters estimated in the eGeMAPS set is available in Appendix A.3. Per Eyben et al. (2013), the parameters can be grouped into the following types: temporal (e.g., speech rate), frequency (e.g., fundamental frequency which is associated with perceived pitch), spectral (e.g., Mel-frequency cepstral coefficients, i.e. MFCCs<sup>5</sup>), and relative energy within specified frequency bands, and energy/amplitude (e.g. intensity).

## Outcome Definition

The resulting data contained a single binary outcome variable indicating whether a given five-second clip was made in response to a likability-focused or competence-focused speaking task. The positive class indicated speech made in response to a competence-focused task and the negative class indicated speech produced in response to a likability-focused task.

## Analysis

A schematic of the complete data handling and analysis pipeline is presented in Figure 1.1.

---

<sup>3</sup> During a pilot phase of the project members of the research team reviewed and rated sample recordings for competence-focused and likability-focused speech. One purpose of that review was to investigate different window sizes to better understand the amount of time required for noticeable acoustic patterns to arise. Recordings were reviewed at window lengths of 1, 2, 5 and 10 seconds. The rationale for selecting the 5-second window used here was that it was sufficient to allow perceptible acoustic patterns to arise but short enough to minimize occurrence of multiple, conflicting patterns across a single window of time.

<sup>4</sup> The eGeMAPS feature set is a set of acoustic features, or properties of speech that were identified by Eyben et al. (2009) for the INTERSPEECH Emotion Challenge that same year. It is a subset of the ComParE features described above. The eGeMAPs feature set is described in detail in Eyben et al. (2010; 2015); the authors also introduce the available Python scripts that can be used to identify and extract the low-level features of speech recordings and functional features that make up the eGeMAPS feature set. The available Python scripts are a part of the open Speech and Music Interpretation by Large Space Extraction (openSMILE) library which makes available scripts that permit identification and extraction of acoustic features of speech and music. Low-level features are those acoustic properties that can be measured directly, such as energy. The functional features are those properties of speech that are derived, such as rate of speech, which is a count of words spoken per unit time. The OpenSMILE Python library is freely available (<https://audeering.github.io/opensmile-Python/>).

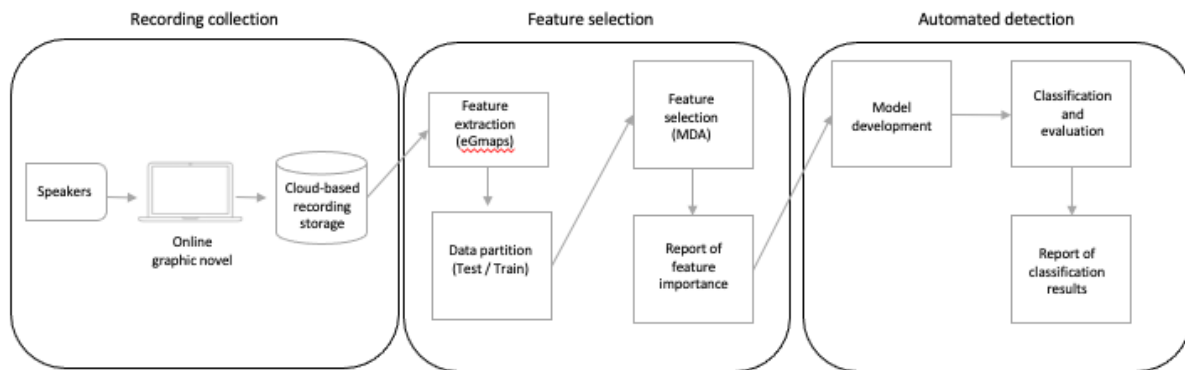
<sup>5</sup> MFCC values indicate measured frequencies of speech sound on a scale that better reflects how speech is perceived by humans (Kent and Read, 1992).

## Data Preparation

After preparation of the digital audio recording files and estimation of the 88 acoustic parameters were carried out, each resulting acoustic parameter was joined with its respective taskID, window rank, recordingID, and unique speakerID. The resulting file was checked for missing values, and no missing values found. All acoustic parameters were standardized, resulting in a mean of 0 and a standard deviation of 1 for each parameter.

**Figure 1.1**

*Schematic of the Complete Data Handling and Analysis Pipeline*



## Data Partitions

Data was partitioned into a training and test set using a 70:30 split. Partitioning was carried out using random selections made at the speaker level to avoid leakage of information between the resulting train and test data sets. The training set represented a total of 1,713 audio clips (821 labeled competence-focused and 892 labeled likability-focused), and the testing set represented a total of 636 audio clips (298 labeled competence-focused and 338 labeled likability-focused)<sup>6</sup>.

## Feature Importance

The importance of each acoustic parameter was estimated through application of a random forest model with the Caret package in R (Kuhn, 2008). Random forests are ensembles of classification, regression, or survival trees (Breiman, 2001). Importantly, random forests make available a set of variable importance measures (VIMs) that provide means for evaluating the importance of a given variable for a classification or regression task relative to the others in the data. Two of the most popular VIMs are the impurity importance and the permutation importance (Breiman, 2001).

The impurity importance is widely referred to as the mean decrease of impurity (MDI), or Gini Importance. It indicates the extent that a given variable in a decision tree results in more homogenous leaves. The permutation importance refers to the mean decrease of accuracy (MDA) that would result if the given variable(s) were removed from the model. In the context of MDA, a variable's importance is larger to the extent that it has a positive effect on the prediction performance of the given model. Impurity importance has been found to be biased in favor of

<sup>6</sup> Data in this and similar studies is necessarily nested with 5-second audio clips nested within longer recordings which are in turn nested within recording tasks and speakers. The fact that the data is structured in this manner is not treated in the modeling stage. Ideally, that structure would be incorporated into the models used. Unfortunately, few multi-level versions of popular machine-learning models are currently available in existing libraries, though this is beginning to change in applications of machine learning in education (Cannistra et al., 2021) and public health (Ji et al., 2020), e.g.

continuous features, factors that have many categories, and features that have categories that occur with high frequency in the data (Nembrini et al., 2018). As a result, permutation importance, or mean decrease in accuracy, is used here.

The process of calculating MDA values using a random forest utilizes permutation of out of bag (OOB) samples to compute the importance of a given variable. OOB samples are observations that were not used in construction of a given tree within a random forest. The collection of OOB observations is used to estimate the prediction error for a given tree and then to evaluate the importance of one or more variables by removing them from the feature set and recalculating the prediction error of the tree (Janitza et al., 2016; Han et al., 2016). For each tree in a random forest, the prediction error (error rate in the case of classification problems) is calculated using the OOB observations. The same calculation is repeated after permuting each feature, or predictor. The differences between the two classification errors – before and after permutation – are averaged over all the trees (Han et al., 2016). Following Janitza et al. (2016) and Han et al. (2016), the equation for the mean difference in accuracy can be specified as follows:

$$MDA_i = \frac{1}{ntree} \sum_{t=1}^{ntree} (EP_{ti} - E_{ti}) \quad (1.1)$$

where:

- $ntree$  indicates the number of trees in the given random forest;
- $E_{ti}$  indicates the OOB error on tree  $t$  before permuting values of feature  $X_i$ ;
- $EP_{ti}$  indicates the OOB error on tree  $t$  after permuting values of feature  $X_i$ .

This same procedure is repeated for all variables, across all trees. Larger MDA values for a given variable indicate its importance for prediction accuracy relative to the other variables used in the random forest model.

### Modeling Approach

#### The L1 Logistic Regression Classifier

L1 logistic regression is used to model the probability of a given audio clip being assigned to a competence-focused or a likability-focused prompt label. The model yields a number between 0 and 1 representing the probability of class membership. In the proposed use, the threshold probability—the probability at which an audio clip has an equal probability of either being a member of the given speech type class or not—is set to 0.5.

Assuming the speech type outcome is denoted as  $Y$ , which has a binary outcome that is 0 if the label is not the targeted speech type and 1 if it is, and assuming the predictors, or features, are denoted as  $X$ , the aim is to model the conditional probability that the outcome  $Y$  has a value of 1 given the predictors  $X$ . This conditional probability is denoted by  $p(Y=1|X)$ . The full logistic regression model can be presented as a regression of the log-odds, so that:

$$\log \left( \frac{p(Y=1|X)}{1-p(Y=1|X)} \right) = \beta_0 + \beta_1 X + \dots + \beta_n X \quad (1.2)$$

where the expression,  $\log \left( \frac{p(Y=1|X)}{1-p(Y=1|X)} \right)$ , is the logarithm of the odds,  $\beta_0$  is the intercept, and  $\beta_1 \dots \beta_n$  describe the weights associated with each of the modeled predictors (or features) of the given audio clip.

In the supervised machine learning context, the objective is to estimate values of  $\beta_0$  and each of the weights  $\beta_1 \dots \beta_n$ , the sum of which results in a probability of  $X$  that most accurately

classifies all the observed data (Hastie et al., 2009; James et al., 2017). Those observations where  $Y$  belongs to the targeted speech type should have a probability as close as possible to 1, and those that do not, should have a probability as close as possible to 0.

Following Hastie et al. (2009), this objective can be rephrased in terms of maximizing the product of these two probabilities, i.e., the likelihood:

$$\log (\prod_{i:Y_i=1} p(X_i) \prod_{j:Y_j=0}(1 - p(X_j))) \quad (1.3)$$

where  $\Pi$  denotes the products over  $i$  and  $j$  which run over the observations classified as 1 and 0 respectively.

Alternatively, one can also rewrite Equation 2 in the form of the *negative* log likelihood:

$$L = -\log (\prod_{i:Y_i=1} p(X_i) \prod_{j:Y_j=0}(1 - p(X_j))) \quad (1.4)$$

in which case the objective is to estimate the intercept,  $\beta_0$ , and the given weights  $\beta_1 \dots \beta_n$ , by minimizing  $L$ .

### Optimization of the L1 Logistic Regression Classifier

L1 logistic regression, or lasso regularization, adds a penalty term,  $\lambda$ , to the log likelihood function:

$$L + \lambda \sum |\beta_1 \dots \beta_n| \quad (1.5)$$

Terms  $\beta_1 \dots \beta_n$  represent features, or measured properties from 1 to  $n$ , and their associated regression weights,  $\beta$ . The term  $\lambda$  is a free parameter, or hyperparameter, with a value that is selected to minimize the error that results when running the eventual model on data comprising the test set, i.e., the out-of-sample error. The lasso accomplishes this by shrinking some of the estimated coefficients, or regression weights, toward or equal to zero. The latter can occur when the penalty is sufficiently large. As a result, the lasso, or L1 regression, is sometimes used to select the variables to be modelled.

Because L1 regression can shrink coefficients to zero, its use can lead to models that are more sparse than standard regression models and may be easier to interpret as a result. In the proposed investigation, the optimal value of  $\lambda$  is estimated through use of grid search with cross-validation<sup>7</sup>, a process that is handled through use of the R library *glmnet* (Friedman et al., 2021). The resulting optimal penalty term,  $\lambda$ , is applied to all weights except for the intercept.

### The Support Vector Machine (SVM)

As noted previously, SVMs have been used with good results by others using acoustic features of speech to infer affect and social signals. In cases in which more than two predictors,

---

<sup>7</sup> The cross-validation framework, or  $k$ -fold cross validation framework, provides a means to test the performance of the machine learned models without use of a new sample of data. It is usually employed in what is referred to as the ‘model selection’ phase of the model development process. The  $k$ -fold cross validation process (KCV) consists of splitting the training data into  $k$  independent subsets, or ‘folds’. All but a subset of the resulting  $k$  folds is used to train the given model and the remaining subset is used to test the model by evaluating the accuracy of the model’s classifications. The training and testing process is repeated until each fold of the data has been used to test the given model and any accompanying hyperparameters. Values for the model’s hyperparameters can be changed and the resulting model tested again by applying the cross-validation process. Through repetition of the cross-validation process multiple hyper-parameters can be evaluated.



or features, are used, the SVM learns from the training instances by mapping them to the feature space and then constructing one or more hyperplanes that separates the instances into two classes, forming a decision boundary (Hastie et al., 2009; James et al., 2017).

A hyperplane is a flat affine subspace with one less dimension than the outcome space in which it is embedded so that, assuming a  $p$ -dimensional space, a hyper plane will have  $p-1$  dimensions (James et al., 2017). As a result, in a two-dimensional space such as a cartesian coordinate system with two axes, the associated hyperplane will be a line. In a three-dimensional space, such as a coordinate system with three axes, the associated hyperplane will be a plane.

Following Hastie et al. (2009), the notion of a decision boundary can be formalized by describing a typical binary classification scenario in which there exists an  $n \times p$  matrix  $X$  comprised of  $n$  observations in  $p$ -dimensional space,

$$\mathbf{x}_1 = \begin{pmatrix} \mathbf{x}_{11} \\ \vdots \\ \mathbf{x}_{1p} \end{pmatrix}, \dots, \mathbf{x}_n = \begin{pmatrix} \mathbf{x}_{n1} \\ \vdots \\ \mathbf{x}_{np} \end{pmatrix}, \quad (1.6)$$

and a set of  $n$  associated outcomes that fall into two classes so that  $y_1, \dots, y_n \in \{-1, 1\}$  where  $-1$  identifies one class and  $1$  identifies the second class.

Classification using a hyperplane assumes it is possible to construct a plane with  $p-1$  dimensions such that the plane separates the training observations according to their respective class labels, in this case  $-1$  and  $1$ . Such a separating hyperplane has the property that on one side of the boundary the class labels have a value of  $-1$ , and on the other side of the boundary they have a value of  $1$ . Again, following the notation of Hastie et al. (2009), in the case of a two-dimensional outcome space, such a hyperplane has the following properties:

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} > 0 \text{ if } y_i = 1 \quad (1.7)$$

and

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} < 0 \text{ if } y_i = -1 \quad (1.8)$$

Where such a hyperplane is possible, it can be used as the basis for a classifier.

Beyond simply identifying the position of an observation relative to the hyperplane, the observation's perpendicular distance from the hyperplane can also provide information about its label. When the magnitude of the perpendicular distance between an observation and the hyperplane is large, the observation is located far away from the hyperplane and one can be more confident about its class assignment. Conversely, when the distance between a hyperplane and a given observation is small, confidence in its associated label is less justified.

Once one or more hyperplanes have been constructed, use of the SVM allows previously unexamined instances to be mapped to the feature space, and their distance from the existing, learned hyperplane(s) can be evaluated. These new instances can then be labeled depending on their position and distance from the hyperplane(s). The distance from the given instance perpendicular to the given hyperplane can be used to inform the certainty of the resulting classification (James et al., 2017).

In the case of the support vector *classifier* (SVC), the resulting hyperplanes are linear (James et al., 2017). A distinguishing feature of support vector machines is that they create a

*non-linear* decision boundary using either a radial kernel or a polynomial kernel with a specified degree. A non-linear support vector machine with a radial kernel is employed here.

### **Optimization of the Support Vector Machine**

The support vector machine presents two parameters that must be tuned to maximize its ability to accurately separate classes of observations in a manner that generalizes to new data. These hyperparameters are cost ( $c$ ) and the hyperparameter  $\gamma$ . When constructing one or more hyperplanes, the location and shape of the hyperplane(s) is determined by optimizing against two competing objectives. On the one hand, generalizability of the SVM can be improved where the distance between the hyperplane(s) and the classes of observations is maximized in the training set. On the other hand, accuracy of the model is improved by maximizing the number of observations that are correctly classified in the training set. This trade-off in generalizability of the model and its accuracy is partially controlled by the value assigned to the cost hyperparameter,  $c$ , which adds a penalty for each misclassified data point.

When the value of  $c$  is small, the associated penalty for misclassifications is also small. This results in larger margins between the hyperplane(s) and classes but also results in a greater number of misclassifications. By contrast, when the value of  $c$  is large, so is the penalty for misclassification of observations. As a result, there are fewer misclassifications, but the margins are also narrower. At the extreme, overfitting can result with large values of  $c$ , and model performance can be expected to decline when run on data other than the training set.

The hyperparameter  $\gamma$  is used with the support vector machine, which specifies non-linear hyperparameters. Informally,  $\gamma$  can be understood to determine the influence of single observations. Large values for  $\gamma$  can result in construction of hyperplanes that are overfit to a small number of observations closely clustered together. On the other hand, values for  $\gamma$  that are very low result in hyperplanes that do not adjust to the complexity of the data and risk underfitting.

As carried out here, optimal values for the cost and  $\gamma$  parameters of the SVM are determined through use of a grid search implemented within a cross-validation framework. This allows empirical discovery of values for the two hyperparameters. Development and estimation of the support vector machines was carried out using the Caret package in R (Kuhn, 2008).

### **Model Evaluation**

All model evaluation metrics are calculated by applying the optimized models to the test set. Per Schuller et al. (2012) as a part of the widely recognized INTERSPEECH Challenges, a set of machine learning challenges that emphasize use of acoustic features of speech, two metrics are used to evaluate the performance of the study's supervised models: unweighted average recall (UAR), and the AUC, the area under the Receiver Operating Characteristic curve (ROC). The motivation for use of the unweighted average recall is that it can be used in settings where there is class imbalance (Schuller et al., 2012; 2013). Motivation for utilizing the AUC also derives from its extensive use in automated detection of social signals and emotion, allowing for comparison of past and current efforts (Schuller et al., 2012).

### **Unweighted Average Recall**

Given two classes of observations,  $X$  and its complement, unweighted average recall<sup>8</sup> can be specified as,

---

<sup>8</sup> As suggested by Equation 8, *recall* is the proportion of true positive classifications made by the model to the sum of its true positive *and* false negative classifications:  $\frac{TP_X}{TP_X + FN_X}$

$$UAR = \frac{1}{2} \left( \frac{TP_X}{TP_X + FN_X} + \frac{TP_{\sim X}}{TP_{\sim X} + FN_{\sim X}} \right) \quad (1.9)$$

where:

- UAR is the unweighted average recall;
- $TP_X$  is the number of accurate classifications of class X made by the model;
- $FN_X$  is the number of false negative classifications of class X made by the model;
- $TP_{\sim X}$  is the number of accurate classifications of the compliment,  $\sim X$ , made by the model;
- $FN_{\sim X}$  is the number of false negative classifications of the compliment made by the model.

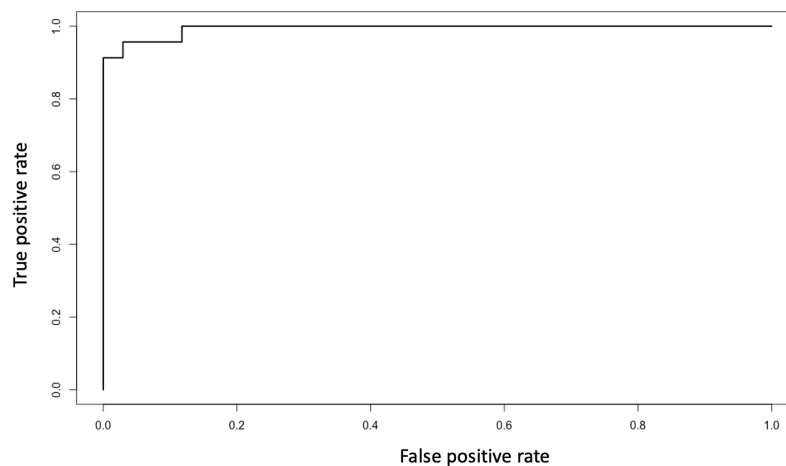
### Area Under the Curve

The AUC, or area under the curve is used for binary classification problems. It is a single value indicating the area under the Receiver Operator Curve (ROC). The ROC is a plot of the true positive rate versus the false positive rate calculated for all threshold values for a model (Hajian-Tilaki, 2013; James et al., 2013). An AUC value of 0.5 indicates a model is performing close to chance. A value of 1 indicates the model is perfectly classifying cases, and a value of 0 indicates it is inverting all classes. A sample receiver operator curve is presented in Figure 1.2, for reference. The ROCR package (Singh et al., 2005) is used to calculate the model performance metrics.

The performance benchmark employed here is presented by Schuller et al. (2012) as part of the INTERSPEECH Challenge for 2012. Using a random forest classifier, Schuller et al. (2012) achieved an unweighted average accuracy of 0.59 with an AUC of 64.7 in a binary classification task classifying speech as likable or not likable. In both cases, the results are close to but better than chance.

### Figure 1.2.

*Sample Receiver Operator Curve*



Note: The true positive rate is also referred to as ‘sensitivity’. True positives are test or model results that correctly identify the presence of a condition or characteristic. The false positive rate is a test or model results that mistakenly identify the presence of a condition or characteristic when it is not present. The true positive rate is the proportion of true positives to the total of true positive results *and* the total of all false negative results:  $TPR = TP / (TP + FN)$ . The false negative rate is the proportion of false negatives to the total number of false negatives *and* true positives:  $FNR = FN / (FN + TP)$ .

**Sensitivity and Specificity**

While not a part of the benchmark metrics, values for the sensitivity and specificity of the models are also provided. Sensitivity is the proportion of instances of the primary class that are correctly identified as such by the model, in this case the primary class is competence-focused speech. The sensitivity of a model ranges from 0 to 1. A value of 0 indicates no primary classes were correctly identified, and a value of 1 indicates all primary classes were correctly identified. Specificity indicates the proportion of secondary classes that are correctly identified as such, in this case audio clips made in response to prompts for likability-speech. Specificity takes on the same range of values as sensitivity. Values of 0 and 1, respectively, also indicate none or all instances of the secondary class were detected.

**Results**

In what follows the top performing acoustic features for distinguishing competence-focused speech and likability-focused speech are described with reference to their estimated MDA values. The study's models are then evaluated for their associated accuracies and AUCs.

**Variable Importance**

Table 1.2 lists the ten acoustic parameters with the highest importance scores among the expert speakers' recordings. Importance estimates for all acoustic parameters are provided in Appendix A.3. The directions of the relations between the various acoustic parameters and the outcome (competence-focused speech and likability-focused speech) are indicated through use of their correlations, as presented in Table 1.3.

**Table 1.2***Descriptions of Features with the Ten Highest Estimated Mean Difference in Accuracy Values*

Feature (Rank)	MDA	Acoustic Property	Category
F5 (1)	7.03	<b>F0semitoneFrom27.5Hz_sma3nz_percentile80.0</b> Percentile 80-th of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz.	Frequency Related (Pitch)
F70 (2)	6.47	<b>mfcc1V_sma3nz_stddevNorm</b> Coefficient of variation of Mel-Frequency Cepstral Coefficient 1 in voiced regions.	Spectral Related (balance)
F40 (3)	5.84	<b>logRelF0H1A3_sma3nz_stddevNorm</b> Coefficient of variation of the ratio of energy of the first F0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3) in voiced regions.	Spectral Related (balance)
F32 (4)	5.33	<b>jitterLocal_sma3nz_stddevNorm</b> Coefficient of variation of the deviations in individual consecutive F0 period lengths.	Frequency Related
F6 (5)	5.21	<b>F0semitoneFrom27.5Hz_sma3nz_pctlrange02</b> Range of 20-th to 80-th of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz.	Frequency Related (Pitch)
F35 (6)	5.19	<b>HNRdBACF_sma3nz_amean</b> Mean harmonics to noise ratio for voiced segments of recordings.	Energy Related
F4 (7)	4.88	<b>F0semitoneFrom27.5Hzsma3nzpercentile50</b> Percentile 50-th of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz.	Frequency Related (Pitch)
F86 (8)	4.59	<b>MeanUnvoicedSegmentLength</b> Mean duration of unvoiced segments.	Temporal Related
F2 (9)	4.46	<b>F0semitoneFrom27.5Hzsma3nzstddev</b> Coefficient of variation of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz.	Frequency Related (Pitch)
F63 (10)	4.10	<b>slopeV0500_sma3nz_amean</b> Mean of linear regression slope of the logarithmic power spectrum within 0-500 Hz band in voiced regions.	Spectral Related (balance)

*Note.* Descriptions derived from Corrales-Astorgano et al., (2018).

It is notable that correlations between the acoustic parameters presented in Table 1.3 and the outcome labels are small. Acoustic features related to the fundamental frequency, or pitch, of the speech sound represent four out of ten of the parameters with the highest MDA values.

**Table 1.3**

*Point Biserial Correlations Between Features with the Highest Mean Difference in Accuracy Estimates and the Outcome, Task Type: Competence-Focused Speech Labeled 0 and Likability-Focused Speech Labeled 1*

Acoustic Feature	Category	Pt. Biserial	Relation	
F0semitoneFrom27.5Hz_sma3nz_percent80.0	Pitch	-0.0890	Competence-Focused Likability-Focused	↑ ↓
mfcc1V_sma3nz_stddevNorm	Spectral	+0.0338	Competence-Focused Likability-Focused	↑ ↓
logRelF0H1A3_sma3nz_stddevNorm	Spectral	-0.0129	Competence-Focused Likability-Focused	↑ ↓
jitterLocal_sma3nz_stddevNorm	Temporal	+0.0555	Competence-Focused Likability-Focused	↑ ↓
F0semitoneFrom27.5Hz_sma3nz_pctlrange02	Pitch	-0.0887	Competence-Focused Likability-Focused	↑ ↓
HNRdBACF_sma3nz_amean	Energy	-0.0498	Competence-Focused Likability-Focused	↑ ↓
F0semitoneFrom27.5Hzsma3nzpercentile50	Pitch	-0.0734	Competence-Focused Likability-Focused	↑ ↓
MeanUnvoicedSegmentLength	Temporal	-0.0469	Competence-Focused Likability-Focused	↑ ↓
F0semitoneFrom27.5Hzsma3nzstddev	Pitch	+0.0800	Competence-Focused Likability-Focused	↑ ↓
slopeV0500_sma3nz_amean	Spectral	-0.0474	Competence-Focused Likability-Focused	↑ ↓

### Model Performance

Here, results from the study's models are discussed in terms of the two planned criteria: unweighted average accuracy (UAR) and AUC. The models' sensitivities and specificities are also given.

**Table 1.4**

*Classification Performance Metrics for Expert Speakers Test Set*

	Sens	Spec	UAR	AUC
L1 Logistic Regression	0.37	0.74	0.56	0.567
Support Vector Machine (Radial Kernel)	0.55	0.55	0.55	0.560

*Note.* Legend: Sens: sensitivity; Spec: specificity; UAR: unweighted average accuracy; AUC: area under the curve.

Resulting performances for both the L1-regression and SVM model are summarized in Table 1.4. When used with speech recordings made by expert speakers in the training data, the optimized L1 logistic regression trained classifier drove coefficients to 0 for all acoustic parameters except for seven. These were F2, F6, F32, F37, F45, F73, and F82, in addition to an

intercept. Applied to the training data, the model had an unweighted average accuracy of 0.58 and an AUC of 0.621 given the task of distinguishing competence-focused speech and likability-focused speech. Sensitivity was 0.37 and specificity was 0.79. When applied to the expert speakers in the test data, the L1 logistic regression-trained classifier had an unweighted average accuracy of 0.56 and an AUC of 0.567. Sensitivity was 0.37 and specificity was 0.74 for the test set.

A linear support vector classifier and support vector machine with a radial kernel were developed and compared. Both models were optimized within a 10-fold cross validation framework. All 88 acoustic parameters were included in the modeling procedure. The support vector machine outperformed the support vector classifier.

When applied to the training data, the optimized support vector machine had an unweighted average accuracy of 0.94 and an AUC of 0.985 when distinguishing between the two types of speech. Sensitivity (true positive rate, TPR) was 0.93 and specificity (true negative rate, TNR) was 0.95. When applied to the test data, the optimized support vector machine had an unweighted average accuracy of 0.55 and an AUC of 0.560. Sensitivity for the model was 0.55 and specificity was 0.55.

## Discussion

### Characteristics of Speech Sounds

The general profile suggested by the acoustic features with the highest MDA values is that higher-pitched speech with longer pauses tends to be associated with likability-focused labels. Speech that exhibits higher variability in pitch and shorter pauses or unvocalized regions tends to be associated with competence-focused labels.

Four of the top parameters used in the models related to speakers' pitch. Correlations between the pitch-related features F4, F5, and F6 and the outcome all indicate that the likelihood that a clip will be identified as competence-focused tends to decrease as the mean pitch of the speech increases. On the other hand, as indicated by the correlation between acoustic parameter F2 and the outcome, as variability in pitch is increased (as indicated by higher standard deviations), the likelihood that speech will be labeled competence-focused increases. In terms of speakers' behavior, higher pitch seems to map onto likability-focused speech, while changes in pitch, which may indicate dynamism, seem to indicate competence-focused speech.

Variability in jitter, parameter F32, is positively related to likability-focused speech and negatively associated with competence-focused speech. Jitter is a measure of variation in the fundamental frequency. Perceptually, jitter has been associated with voice roughness (Rabinov et al., 1995). The harmonic to noise ratio (HNR) indicates the ratio of harmonic and nonharmonic components of the speech sound, measured in decibels. Lower values of HNR indicate more noise in the speech sound. In the case of the current study, lower mean HNR values are also associated with competence-focused labels.

### Model Results

The model results for the current study are lower than those achieved by Schuller et al. (2012) in their 2012 INTERSPEECH Challenge. In the case of Schuller et al. (2012), labels were assigned by a team of raters. Speech was scripted with single sentence utterances and collected by phone in naturalistic environments. The difference in performance between the models presented here and those of Schuller et al. (2012) two types of labels and recording scenarios may be of interest to those researchers wanting to apply automated detectors in real-world environments and contexts, such as educational contexts which are highly variable and do not afford tight controls on environmental and speaker variables.

### **Conclusion**

A machine learning approach utilizing L1 logistic regression and a support vector machine with a radial kernel was used for the binary classification task of distinguishing competence-focused and likability focused speech. Recordings were gathered from expert speakers who provided extemporaneous responses to recording prompts in uncontrolled environments. The importance of a subset of acoustic features of speech sounds for the classification of speech was investigated. While performances of the resulting classification models used were similar to a recognized benchmark for a similar classification task, their performances were slightly lower. The models and online recording platform are made publicly available with the suggestion that they may be used to further improve automated detection of social stance from extemporaneous speech in uncontrolled environments.



## CHAPTER TWO

### Detection of Social Signals Among Non-Expert Speakers

In the previous chapter, the emphasis was placed on detection of social stance through acoustic features of speech. Communication of a social stance can be understood as involving sets of *social signals*. Social signals are associated with behaviors that individuals use to establish and/or maintain bonds with others (Dunbar, 1996). Functionally, individuals often use social signals to solve a wide range of problems that arise while initiating and maintaining coordinated action. While many social signals rely on linguistic or lexical information, others utilize combinations of gestures, interactions with material objects, and acoustic features of speech. In the case of the latter, prosodic features of speech—loosely characterized as the changes in the pitch, intensity, duration and rate of speech—have been recognized as producing what can be referred to as the song of language (Boustien, 2003; Patel & Daniele, 2003; Palmer & Hutchins, 2006). These characteristics have also been recognized for their important role in conveying social signals (Pentland, 2004; Vinciarelli et al., 2009). Viewing social stances as a product of one or more social signals delivered via speech or other modes of communication connects the work presented in Chapter 1 to new literature and new constructs such as ethology, behavioral repertoires, animal communication, language development, and social perception.

If automated systems are to detect social signals in everyday, or what may be called ‘naturalistic’ settings, they will need to be capable of utilizing prosodic features of natural, unscripted speech that is produced by non-experts in uncontrolled environments. The current study extends work described in Chapter 1 by utilizing prosodic features of speech to automate detection of extemporaneous competence-focused and likability-focused speech from *non-expert* speakers in uncontrolled environments.

Recordings from a new sample of speakers are used for the current study. All participants are judged to be non-experts in the sense that they did not report training or experience in acting or voice-over work, and all reported that they are native English speakers living within the continental United States. Making use of the new data set, machine learning models were developed using L1 logistic regression, a support vector classifier, and a support vector machine using a radial kernel.

The models’ performance was evaluated for accuracy in differentiating between speech associated with likability-focused and competence-focused recording tasks. Acoustic features of speech that were identified as contributing information to the classification of likability-focused and competence-focused speech among speakers were used to describe and compare differences in speaker behavior. Existing benchmarks for performance from Schuller et al. (2012; 2013) and Ranganathan et al. (2013) were utilized for comparison. Schuller et al. (2012) achieved a baseline unweighted average recall of 0.590, and an AUC of 0.647 when classifying likability speech from acoustic features. While these results have served as a benchmark for a great deal of work in this area, it is notable that they are not far from chance<sup>9</sup>. This is taken to indicate the difficulty of the task of automating detection of social signals from acoustic features of speech as well as the need for additional work in this area.

---

<sup>9</sup> An AUC of 0.5 indicates performance at the same level of chance. An AUC value of 1.0 indicates perfect classification and a value of 0 indicates failure to successfully classify any cases. Both unweighted average recall and AUC are described in more detail in the methods section.

### **Purpose**

This study extends the investigation described in Chapter 1 by focusing on prosodic features of speech and the use of non-expert speakers who generated extemporaneous speech in uncontrolled environments. With these two additions, the purpose of this study is two-fold. First Chapter 2 investigates the feasibility of using prosodic features of speech to automate detection of social signals from speech generated by non-experts speaking extemporaneously within naturalistic environments. It accomplishes this through the development and evaluation of a set of machine learning models capable of classifying the study's new speech samples as either competence-focused or likability-focused. The models' performances are subsequently evaluated. Second, Chapter 2 provides descriptions of the prosodic features of speech that are productive in distinguishing likability-focused speech from competence-focused speech. During the course of the current paper, it is recognized there may be a limit to how accurately systems can be when identifying social signals when only a single source of information is used—e.g., prosodic features of speech. This sets the stage for work presented in Chapter 3, which investigates detection of social signals across multiple sources of information.

### **Intended Contributions**

As mentioned above, automated detection of social signals from prosodic features of speech is a challenging task. Use of extemporaneous speech from non-expert speakers in naturalistic environments adds to that challenge. But employing such conditions ultimately stands to make the resulting models more generalizable to a broader range of language communities and contexts than they would be if utilizing more contrived conditions, e.g. scripted speech recorded in clinical settings by expert speakers. Ultimately this study aims to contribute to ongoing work to automate detection of social signals by better approximating conditions *in-the-wild*, or training the study's models through examples of extemporaneous speech from non-experts in naturally occurring environments. Further, by focusing on prosodic features of speech the study also aims to contribute to an understanding of behaviors speakers may be able to learn and adopt to better convey likability and competence. Lastly, evidence is presented throughout that supports the claim that communication of social signals likely occurs across multiple communication channels involving acoustic and lexical features of speech, in addition to gestures, facial expressions, and interactions with material objects. While methods for detecting social signals from acoustic features of speech such as prosody can certainly be improved, it is likely that automated detection will require systems capable of utilizing information from multiple channels. This is necessarily true where information is either non-redundant across channels or interactions between information from multiple channels enhance, modulate, or otherwise change the message and/or response (Partan & Marler, 1999; 2005). While only prosodic features of speech are utilized in the current chapter, considerations such as these set the stage for Chapter 3, which investigates the use of speech as a *composite signal*, containing both lexical information (*what was said*) in addition to acoustic features of speech (*how it was said*).

### **Background**

In the following section, the roles of prosodic features of speech in conveying social signals are described. Their importance in initiating, maintaining and terminating coordinated action is emphasized. Interestingly, prosodic features of speech are often used in concert with the propositional content of speech as well as gesture and facial expression. While this is not a theme treated in depth in the current chapter, it is an important theme for Chapter 3.

### **Social Signals and Prosodic Features of Speech**

Social signals are associated with behaviors that individuals use to establish and/or maintain bonds with others (Dunbar, 1996). Social signals express attitudes toward individuals, groups, activities, and social situations, and they are manifested through cues that include facial expressions, body postures, gestures, and paralinguistic aspects of speech (Vinciarelli et al., 2009). Among the functions that social signals play is the prevention and solution of challenges that arise with coordinated action (Nesse, 2007; Barclay, 2011; Dessales, 2014).

Successfully initiating and completing coordinated action requires two or more individuals “acting-together” across “thin slices of time” to (Noe, 2006; Taborsky, 2007). This requirement for coordinated action underscores the fact that individuals on the brink of, or who are actively engaged in cooperating, often must coordinate their action(s) in real time requiring them to express and interpret communications in fractions of a second. Though instances of cooperation have been documented in which there is little or even no attention paid to the status of the cooperating other and their contributions (Enders & Ward, 1985; Wickler & Seibt, 1993), these cases are taken to be an exception. Typically, acting-together is understood to require regular communications between actors in order to solve what may be termed ‘coordination problems.’ These coordination problems include development and confirmation of joint attention, communication of sustained commitment, and communication of shared intention (Tomasello & Carpenter, 2007).

While the linguistic, or propositional, content of speech no doubt plays an important role in solving such coordination problems, evidence also points to an important role for prosodic features of speech, such as changes in pitch and loudness. A growing body of evidence indicates that social actors use social signals that are communicated through use of prosodic features of speech to solve everyday coordination problems. As social actors face challenges associated with cooperating with others, they manipulate prosodic features of their speech in different ways to communicate a broad range of social signals as a means of soliciting, initiating, maintaining, and terminating cooperative activities.

### **Solving Coordination Problems in Small Windows of Time**

Growing bodies of theory and research now point to actors’ use of prosody to solve a wide range of coordination challenges. Referring to its *guiding* function, for example, Darwin (1975) has proposed that prosodic features of speech are used by speakers to solve coordination challenges related to generating shared, or joint, attention. Darwin’s assertion is bolstered by growing evidence that speakers utilize prosody to focus the attention of their interlocutors. Gupta et al. (2012) show that manipulation of prosodic features of speech such as intensity and pitch predict joint attention and engagement in children. Barry (1981) found that prosodic features of speech alone were sufficient to focus interlocutors’ attention on new and desired information.

Prosody’s role in conveying a speaker’s emotions also aids in creating shared, or joint, attention. Brosch et al. (2008), as well as Rigoulot and Pell (2012) and Paulmann et al. (2012), for example, have all found evidence that speakers used prosody to guide interlocutors’ attention to specific features, objects, or agents in the environment. Prosody can be coordinated with gesture to have similar effects. Research in this area describes how prosodic features of speech that are tightly coordinated with simple repetitive gestures such as pointing and rhythmic beating or pounding movements—*prosodic temporal alignment*—also aid joint attention (Bull & Connelly, 1985; Hadar et al., 1984; Hadar et al., 1983, 1984; Kendon, 1994; Jesse & Johnson, 2012). More specifically, when coordinated with changes in speakers’ pitch (Feyereise, et al., 1988; Cave et al., 1996; McNeill, 2000; Yehia et al., 2002), or with percussive, rhythmic emphases in

speech (Freedman & Hoffman, 1967; Ekman & Friesen, 1969), gestures can direct the eye gaze of interlocutors and resolve ambiguous referents.

Finally, prosodic features of speech have also been found to play a role in communicating one's commitment to the propositional content of utterances and one's commitment to action. Heritage (2013), Borrás-Comes et al. (2011), and Swerts and Kraemer (2005) for example, have all found that prosodic features of speech, such as slow speech rate and rising intonation that occur in fractions of a second at the end of statements can be used to predict speakers' confidence in their claims. Likewise, researchers have found that speakers' changes in pitch and rate of speech may be used by interlocutors to form judgements of speakers' sincerity (Haiman, 1998; Attardo et al., 2003; Cheang & Pell, 2008), and to detect deception (Levitan, Maredia and Hirschberg, 2018; Chen, Levitan, Levine et al., 2020), agreement, and commitment (Bousmalis et al., 2009; Bousmalis et al., 2011). These findings are reproduced across species. For example, non-human primates have also been found to use vocal cues to signal readiness and commitment to act (Nesse, 2002; Silk, 2002).

### **Solving Coordination Problems Across Larger Windows of Time**

Relevant interactions also take place within a broader social context that can develop more slowly, and that presents a different set of coordination challenges. One such challenge arises with the potential of *free-riders*, or individuals that benefit from coordinated action without fairly contributing. Cooperation with others often delays benefits associated with action and requires investments with uncertain returns (Brosnan et al., 2010). The prospect of delayed benefits and uncertain returns poses problems for potential cooperators as they face the possibility of encountering free riders. In that context, individuals able to convincingly communicate a willingness and ability to affiliate (likability) and ability to successfully carry out joint action (competence) may be at an advantage (Eisenbruch & Krasnow, 2019; Bor, 2020). Likewise, individuals able to accurately interpret such signals and assess their veracity may be able to minimize impacts of hostile and insincere actors, thereby improving their own fitness (Hack et al., 2013). Here too, use of prosody to generate convincing social signals likely plays a key role.

Emphasis on the role of social signaling for successful coordination serves to further underscore the importance of impression management that was mentioned in Chapter 1; the importance of effectively conveying and interpreting social signals, and the potential benefits of socially aware systems that can automatically detect social stances such as likability and competence. While much of the existing work in this area utilizes scripted speech in clinical or research environments, automated systems will most likely need to operate *in-the-wild*—interacting with non-expert speakers who are behaving extemporaneously in naturalistic environments.

### **Multimodal Communication**

Given the importance of social signals such as those involved in conveying competence and likability, it seems reasonable to assume that where possible, such signals are conveyed along multiple channels, or modalities. Communication involving multiple modalities can be referred to as *multimodal communication*, defined as communication occurring through more than one sensory channel (Partan & Marler, 2005). Potential benefits accrue from use of multimodal signaling. In cases in which the signal components provide redundant information, their use can increase the likelihood that the message will be received because redundancy may reduce errors in detection and interpretation (Partan & Marler, 1999; 2005). Use of nonredundant signals can increase the amount and variety of information sent. Partan and Marler's (1999) framework for multimodal signals, presented in Figure 2.1, points to the varied outcomes of multimodal signals.

**Figure 2.1**

*Framework for Classification of Multimodal Signals From Partan and Marler (1999)*

SEPARATE COMPONENTS	MULTIMODAL COMPOSITE SIGNAL
<p><b>Redundancy</b></p> <p>signal      response</p> <p>a      →      □</p> <p>b      →      □</p>	<p>signal      response</p> <p><u>a+b</u>      →      □      <b>Equivalence (intensity unchanged)</b></p> <p><u>a+b</u>      →      □      <b>Enhancement (intensity increased)</b></p>
<p><b>Nonredundancy</b></p> <p>a      →      □</p> <p>b      →      ○</p>	<p><u>a+b</u>      →      □ and ○      <b>Independence</b></p> <p><u>a+b</u>      →      □      <b>Dominance</b></p> <p><u>a+b</u>      →      □ or □      <b>Modulation</b></p> <p><u>a+b</u>      →      △      <b>Emergence</b></p>

*Note.* Redundant signals are represented in the first row and nonredundant signals are represented in the second row. The geometric shapes represent responses to separate components of multimodal signals. Two components of multimodal signals, a and b, are represented though there can be more than two in a given communication effort. In the upper left-hand cell, the components of redundant signals are followed by the same response from an interlocutor. Responses to non-redundant components are not equivalent (represented as different shapes), and the possibility of no response is realistic as well. In the presence of redundancy, the response can either be the same or similar but enhanced. Four possible relations between non-redundant signal components are hypothesized: signal-components engender different and noninteracting responses (independence), one signal component dominates another (dominance), one signal-component modulates another (modulation), or signal-components interact to engender a qualitatively different response (emergence). See Partan & Marler (1999; 2005) for more detailed accounts.

For cases in which signals are non-redundant, Partan and Marler’s (1999) framework describes four possible outcomes: *independence*, *dominance*, *modulation*, and *emergence*. In the case of independence, nonredundant multimodal signals engender a composite response that combines behaviors associated with each of the individual signals. In the case of dominance, the interlocutor’s response occurs as if only one of the signals was present. The case of modulation is similar to dominance, but the interlocutor’s response is amplified. Finally, in the case of modulation, the signals interact in some way to engender a qualitatively different behavior from those engendered by either of the signals or set of signals operating on their own.

Partan and Marler’s (1999) framework for multimodal communication is particularly relevant for those interested in automated detection of social signals, as all the categories of response to nonredundant signals indicate potential sources of failure for such systems if they are unimodal, or only account for a single modality of communication. In cases in which multimodal signals are nonredundant and independent, for example, automated systems may interpret and respond to a single signal or signal set, but not to another, with the result that their response is accurate but incomplete. Similar scenarios arise in the case of modulation in which a more

intense or amplified response is expected but not delivered; and in the case of emergence, the system's inferences and responses to the multimodal signals may be incorrect all together.

Consideration of Partan and Marler's (1999) framework is important for the current study as well because it suggests that there may be a limit to the accuracy of any given unimodal system, such as a system that relies solely on the prosodic content of speech. Though existing methods of automating detection of social signals can certainly be improved, there may be a limit to their accuracy in cases in which only a single communication channel is accounted for, or where an unproductive combination of communication channels is used.

Nevertheless, development of unimodal detectors can still be productive for at least three reasons. First, investigation and development into unimodal detectors can lead to incremental improvements that maximize their accuracy and can be leveraged when several unimodal detectors are integrated to create a multimodal indicator. Second, over the long run, continued efforts to develop and improve unimodal detectors for social signals may eventually inform estimates of the ceiling, or limit, of their performance. This would ultimately contribute to an estimate of the proportion of information conveyed along each channel, or modality on its own. Lastly, in spite of their limitations, unimodal detectors can still have applications in everyday social situations. For example, it is not uncommon to encounter scenarios in which communication signals can be effectively transmitted through only a single modality—radio transmissions, noise filled, or low-light environments may be examples here.

In what follows, the feasibility of automating the detection of social signals with extemporaneous speech from non-expert speakers in naturalistic environments is investigated through use of prosodic features of speech. Prosodic features that best distinguish competence-focused and likability-focused speech under these conditions are identified and described. Throughout, it is recognized that there may be a limit to how well social signals can be detected from a single source of information.

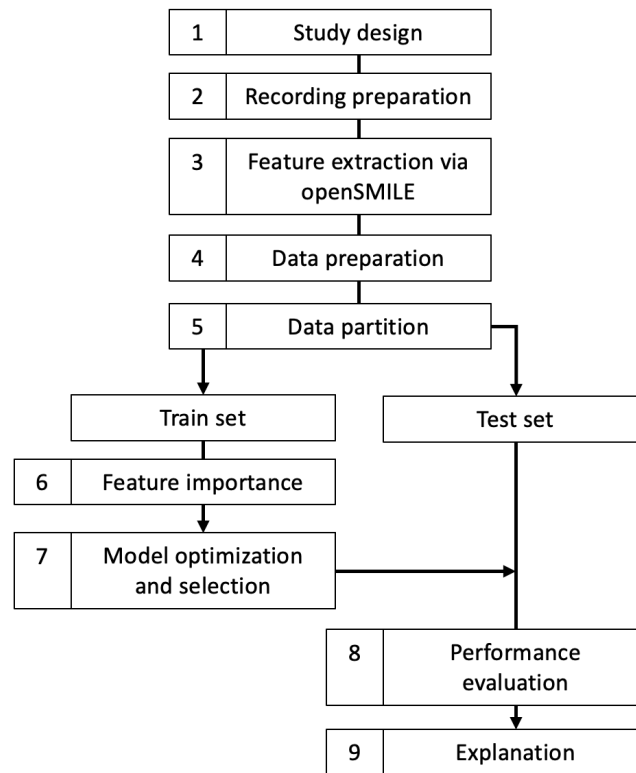
### **Materials and Methods**

The source code for analyses is provided in Appendices A-D in support of transparency and openness. A summary of the data handling and modeling pipeline for the study is presented in Figure 2.2 and discussed in detail below.

#### **Study Design**

This study uses a retrospective cross-sectional design to investigate automated detection of likability-focused and competence-focused speech from non-expert speakers who respond extemporaneously to prompts in naturalistic, uncontrolled environments. As in the study presented in Chapter 1, the aim of this study is to create and compare a set of models for classifying audio recordings of human speech as either likability-focused or competence-focused. Prosodic features of speech are the single source of information used in the study's models.

**Figure 2.2**  
*Data Handling and Modeling Pipeline for the Study*



## Participants

Nonexpert speakers were recruited using the crowdsourcing website Amazon Mechanical Turk.<sup>10</sup> All nonexpert speakers were volunteers and consented to participate in the study. The MTurk setup option was used to limit participation in the study to those workers who had achieved the formal status of high-performing workers, further recognized as masters within the MTurk system (Peer et al., 2014). Additionally, a geographic restriction was implemented, permitting only individuals confirmed as current residents of the United States to participate in the study. Candidate nonexpert speaker participants responded to a demographic survey prior to being accepted in the study. Only those indicating they were native English speakers of American English were selected for participation in order to reduce variability in speech with sources other than the intended speech type.

A total of 154 adults participated in the study. Of those reporting their demographic information, 67 (43.5%) were female and 87 (56.5%) were male, all 18-55 years of age ( $M = 23.24$ ,  $SD = 3.94$ ). All speaker participants were native speakers of American English currently residing in the United States. A summary of participants' reported sex at birth and racial/ethnic backgrounds is provided in Table 2.1.

<sup>10</sup> MTurk; <http://www.mturk.com>.

**Table 2.1**  
*Reported Demographics of Non-Expert Speaker Participants*

	American Indian		Asian		Black		Pacific Islander		White		Multi-ethnic		Decline / Other		Total	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
All	4	2.6	24	15.6	29	18.8	2	1.3	82	53.3	8	5.2	5	3.3	154	100.0
<b>Female</b>	<b>2</b>	<b>1.3</b>	<b>10</b>	<b>6.5</b>	<b>12</b>	<b>7.8</b>	<b>1</b>	<b>0.7</b>	<b>36</b>	<b>23.4</b>	<b>6</b>	<b>3.9</b>	<b>0</b>	<b>0.0</b>	<b>67</b>	<b>43.5</b>
Hispanic	1	0.7	1	0.7	1	0.7	1	0.7	9	5.8	0	0.0	0	0.0	13	8.4
Non-Hispanic	1	0.7	9	5.8	17	11.0	0	0.0	64	41.6	6	3.9	0	0.0	97	63.0
<b>Male</b>	<b>2</b>	<b>1.3</b>	<b>14</b>	<b>9.1</b>	<b>17</b>	<b>11.0</b>	<b>1</b>	<b>0.7</b>	<b>46</b>	<b>29.9</b>	<b>2</b>	<b>1.3</b>	<b>5</b>	<b>3.3</b>	<b>87</b>	<b>56.5</b>
Hispanic	0	0.0	0	0.0	2	1.3	0	0.0	11	7.1	0	0.0	5	3.3	18	11.7
Non-Hispanic	2	1.3	14	9.1	15	9.7	1	0.7	35	22.7	2	1.3	0	0.0	69	44.8

While somewhat balanced regarding sex at birth, it is notable that the sample used for the present study is not balanced with regard to its other demographics. This is relevant for the current study as it likely limits the generalizability of its findings. While different language communities may exhibit similar prosodic patterns in their speech, it is reasonable to also assume variability between groups as well.

### Speech Recording

As in the study presented in Chapter 1, participants were given access to the OnlineRecording System (ORS), a cloud-based authoring tool for graphic novel experiences which integrates audio recording functionality. The ORS system requests access to the microphone on the user's computer and leads them through a series of checks to ensure the microphone is working correctly and the noise-level in the recording environment is acceptable. Users are then introduced to the story line of a graphic novel, titled *Advice Hour*, designed by the author to meet requirements of the current study. Participants are invited to assume the role of a podcast host responding to callers' questions about how to handle specific communication dilemmas in their personal and professional lives. The structure of the *Advice Hour* storyline allows for presentation of multiple caller scenarios, each of which present a different context and explicit prompts for either competence-focused or likability-focused speech. The graphic novel also allows for randomized presentation of the caller scenarios to participants. Each of the scenarios ends with a request prompting speaker-participants to record *what* they would say and *how* they would say it, with explicit instructions to emphasize either competence-focused speech or likability-focused speech. Each of the graphic novel scenarios and recording prompts were tested and iterated using a concurrent think-aloud protocol in order to ensure the prompts for competence-focused and likability-focused speech were understood by speaker-participants as intended. As a part of the ORS functionality, after each response, participants are given the chance to review their recording and either accept or revise it. Approximately twenty percent of participants utilized the review and revise functions of the ORS. The ORS stores participant-accepted recordings in a password protected cloud-based file.



### Speech Pre-Processing

Speaker-participants' recordings were reviewed for evidence of on-task performances. Recordings for each task were then segmented into 5 second clips<sup>11</sup>, each of which was associated with a unique identifier indexing the task, task type (likability-focused task vs. competence-focused task), and window rank of the recording (the cardinal value denoting the position of the clip in the full recording, the first five-second window, second 5-second window, etc.).

### Feature Extraction via openSMILE

Using the OpenSMILE feature library, the INTERSPEECH 2013 Computational Paralinguistics feature set (ComParE) was extracted from participants' five-second clips of speech for each task (Eyben et al., 2013).<sup>12</sup> The ComParE feature set comprises suprasegmental features obtained by calculating a set of functional statistics (mean, range, quartiles, standard deviation, minima and maxima, *inter alia*) to a smaller number of low-level descriptors. The categories of low-level descriptors given in the ComParE feature set are presented in Table 2.2. These include prosodic and spectral features of the speech sound in addition to measures of sound quality. The list of functionals that are subsequently applied to the low-level descriptors are given in Table 2.3.

In total, the ComParE feature set includes 6,347 features that can be categorized as prosodic, spectral, cepstral, or relating to sound quality. Of these, 3,083 relate to prosodic features, all of which are utilized in the current study.

---

<sup>11</sup> During a pilot effort, members of the research team who were engaged with the project reviewed a sample of practice recordings from participants to better understand what window size(s) was appropriate for the study presented here. Windows of size 1s, 2s, 5s, and 10s were investigated during a pilot leading up to the study presented here. Reviewers deemed that 2s windows were too short to allow clear, perceptible patterns to arise in speakers' speech. By contrast, reviewers believed they were able to identify multiple perceptible patterns in 10s windows, some of which suggested contradictory evidence. The 5s window size was judged to most consistently allow enough time for perceptible patterns to arise while minimizing the number of instances of multiple, conflicting patterns within a given clip.

<sup>12</sup> The ComParE feature set is a set of acoustic features, or properties of speech, that were identified for the 2011 INTERSPEECH Challenge. The feature set is described in detail in Eyben et al. (2010); the authors also introduce the available Python scripts that can be used to identify and extract low-level features of speech recordings and functional features that result from application of one or more mathematical functions to the low-level features. The available Python scripts are a part of the open Speech and Music Interpretation by Large Space Extraction (openSMILE) library, which makes available scripts that permit identification and extraction of acoustic features of speech and music. Low-level features are those acoustic properties that can be measured directly, such as loudness. The functional features are those properties of speech that are derived, such as rate of speech, which is a count of words spoken per unit time. The OpenSMILE Python library is freely available (<https://audeering.github.io/opensmile-Python/>).

**Table 2.2***Low-Level Descriptors of the ComParE Acoustic Feature Set (Eyben et al., 2013)*

	Group
<b>Energy Related Low-Level Descriptors</b>	
Sum of auditory spectrum (loudness)	Prosodic
Sum of RASTA-style filtered auditory spectrum	Prosodic
RMS energy, zero-crossing rate	Prosodic
<b>Spectral Low-Level Descriptors</b>	
RASTA-style auditory spectrum, bands 1-26 (0-8kHz)	Spectral
MFCC 1-14	Cepstral
Spectral energy 250-650 Hz, 1-4 kHz	Spectral
Spectral roll-off point 0.25, 0.50, 0.75, 0.90	Spectral
Spectral flux, centroid, entropy, slope	Spectral
Psychoacoustic sharpness, harmonicity	Spectral
Spectral variance, skewness, kurtosis	Spectral
<b>Voicing Related Low-Level Descriptors</b>	
$F_0$ (SHS and Viterbi smoothing)	Prosodic
Probability of voicing	Sound quality
Log. Harmonic-to-noise ratio, Jitter, Shimmer	Sound quality

**Table 2.3***Functionals in the ComParE Feature Set (Eyben et al., 2013).*

	Group
<b>Functionals Applied to LLDs &amp; ALLDs</b>	
Quartiles 1-3, 3 inter-quartile ranges	Percentiles
1% Percentile (~min), 99% percentile (~max)	Percentiles
Percentile range 1-99%	Percentiles
Position of min/max, range (max-min)	Temporal
Arithmetic mean, root quadratic mean	Moments
Contour centroid, flatness	Temporal
Standard deviation, skewness, kurtosis	Moments
Rel. duration LLD is above 25/50/75/90) range	Temporal
Rel. duration LLD is rising	Temporal
Rel. duration LLD has positive curvature	Temporal
Gain of linear prediction (LP), LP coefficients 1-5	Modulation
Mean, max, min, SD of segment length	Temporal
<b>Functionals Applied to LLDs Only</b>	
Mean value of peaks	Peaks
Mean value of peaks – arithmetic mean	Peaks
Mean/SD of inter peak distances	Peaks
Amplitude mean of peaks, of minima	Peaks
Amplitude range of peaks	Peaks
Mean/SD of rising/falling slopes	Peaks
Linear regression slope, offset, quadratic error	Regression
Quadratic regression a, b, offset, quadratic error	Regression
Percentage of non-zero frames	Temporal

**Outcome Definition**

Each audio clip is tagged with a single binary outcome variable indicating whether it was recorded in response to a competence-focused or likability-focused speaking task. The primary, or positive class label indicates speech made in response to a competence-focused task and the secondary class label indicated speech produced in response to a likability-focused task. The classification models presented in the current study will use prosodic features to classify each audio clip according to its competence-focus or likability-focus label.

## Analysis

### Data Preparation

A schematic of the complete data handling and analysis pipeline is presented in Figure 2. After preparation of the digital audio recording files and extraction of the full set of the ComParE acoustic parameters via openSMILE, each resulting acoustic parameter was joined with its respective taskID, window rankID (the cardinal value denoting the position of the clip in the full recording—first 5s window, second 5s window, etc.), recordingID, and unique speakerID as part of the data preparation process. As a part of the data preparation process, the file was also checked for missing values, and no missing values were found. All acoustic parameters were standardized resulting in a mean of 0 and a standard deviation of 1 for each.

### Data Partition

Data was partitioned into training and test sets using a 70:30 train-test split, with random selections made at the speaker level to avoid leakage of information between the resulting train and test data sets. The training set represents a total of 2,413 audio clips, 1,030 (42.7%) labeled competence-focused and 1,383 (57.3%) labeled likability-focused.<sup>13</sup> The testing set represents a total of 968 audio clips, 397 (41.0%) labeled competence-focused and 575 (59.4%) labeled likability-focused.

### Feature Importance

As in the previous chapter, the importance of each prosodic feature is inferred via estimates of the mean decrease in accuracy (MDA) resulting from removal of each feature from a random forest trained on the training data. Random forests are ensembles of classification, regression, or survival trees (Breiman, 2001). Mean decrease in accuracy values for each acoustic parameter were estimated using the Caret package in R (Kuhn, 2008).

The process of calculating MDA values using a random forest includes the permutation of out of bag (OOB) samples to compute the importance of a given variable. OOB samples are observations that were not used in construction of a given tree within a random forest. The collection of OOB observations is used to estimate the prediction error for a given tree and then to evaluate the importance of one or more variables by removing them from the feature set and recalculating the prediction error of the tree (Janitza et al., 2016; Han et al., 2016). For each tree in a random forest, the prediction error (error rate in the case of classification problems) is calculated using the OOB observations. The same calculation is repeated after permuting each feature, or predictor. The differences between the two classification errors—before and after permutation—are averaged over all the trees (Han et al., 2016). Following Janitza et al. (2016) and Han et al. (2016), the equation for the mean difference in accuracy can be specified as follows:

$$MDA_i = \frac{1}{ntree} \sum_{t=1}^{ntree} (EP_{ti} - E_{ti}) \quad (2.1)$$

where:

- *ntree* indicates the number of trees in the given random forest

---

<sup>13</sup> Data in this and similar studies necessarily nested with five-second audio clips nested within longer recordings which are in turn nested within recording tasks and speakers. The fact that the data is structured in this manner is not treated in the modeling stage. Ideally, that structure would be incorporated into the models used. Unfortunately, few multi-level versions of popular machine-learning models are currently available in existing libraries, though this is beginning to change in applications of machine learning in education (Cannistra et al., 2021) and public health (Ji et al., 2020).

- $E_{ti}$  indicates the OOB error on tree  $t$  before permuting values of feature  $X_i$
- $EP_{ti}$  indicates the OOB error on tree  $t$  after permuting values of feature  $X_i$

This same procedure is repeated for all variables across all trees. Larger MDA values for a given variable indicate its importance for prediction accuracy relative to the other variables used in the random forest model.

### Modeling Approach

The modeling approach used here is consistent with the approach specified in the previous chapter: the L1 logistic regression classifier, the support vector classifier, and the support vector machine with radial kernel. Descriptions for each are provided again here for convenience.

### The L1 Logistic Regression Classifier

L1 logistic regression is used to model the probability of a given audio clip being assigned to a competence-focused or a likability-focused prompt label. The model yields a number between 0 and 1 representing the probability of class membership. In the proposed use, the threshold probability, i.e. the probability at which an audio clip has an equal probability of either being a member of the given speech type class or not, is set to 0.5.

Assuming the speech type outcome is denoted as  $Y$ , which has a binary outcome of 0 if the label is not the targeted speech type and 1 if it is, and assuming the predictors, or features, are denoted as  $X$ , the aim is to model the conditional probability that the outcome  $Y$  has a value of 1 given the predictors  $X$ . This conditional probability is denoted by  $p(Y=1|X)$ . The full logistic regression model can be presented as a regression of the log-odds, so that:

$$\log\left(\frac{p(Y=1|X)}{1-p(Y=1|X)}\right) = \beta_0 + \beta_1 X + \dots + \beta_n X \quad (2)$$

where the expression  $\log\left(\frac{p(Y=1|X)}{1-p(Y=1|X)}\right)$  is the logarithm of the odds,  $\beta_0$  is the intercept, and  $\beta_1 \dots \beta_n$  describe the weights associated with each of the modeled predictors (or features) of the given audio clip.

In the supervised machine learning context, the objective is to estimate values of  $\beta_0$  and each of the weights  $\beta_1 \dots \beta_n$ , the sum of which results in a probability of  $X$  that most accurately classifies all the observed data (Hastie et al., 2009; James et al., 2017). Those observations where  $Y$  belongs to the targeted speech type should have a probability as close as possible to 1, and those that do not should have a probability as close as possible to 0.

Following Hastie et al. (2009), this objective can be rephrased in terms of maximizing the product of these two probabilities, i.e., the likelihood:

$$\log\left(\prod_{i:Y_i=1} p(X_i) \prod_{j:Y_j=0} (1 - p(X_j))\right) \quad (3)$$

where  $\Pi$  denotes the products over  $i$  and  $j$  which run over the observations classified as 1 and 0 respectively.

Alternatively, one can also rewrite Equation 4 in the form of the *negative* log likelihood:

$$L = -\log\left(\prod_{i:Y_i=1} p(X_i) \prod_{j:Y_j=0} (1 - p(X_j))\right) \quad (4)$$

In this case, the objective is to estimate the intercept,  $\beta_0$ , and the given weights  $\beta_1 \dots \beta_n$ , by minimizing  $L$ .

### Optimization of the L1 Logistic Regression Classifier

L1 logistic regression, or lasso regularization, adds a penalty term,  $\lambda$ , to the log likelihood function:

$$L + \lambda \sum |\beta_1 \dots \beta_n| \quad (2.5)$$

Terms  $\beta_1 \dots \beta_n$  represent features, or measured properties, from 1 to  $n$ , and their associated regression weights,  $\beta$ . The term  $\lambda$  is a free parameter, or hyperparameter, with a value that is selected to minimize the error that results when running the eventual model on data comprising the test set, i.e., the out-of-sample error. The lasso accomplishes this by shrinking some of the estimated coefficients, or regression weights, toward or equal to zero. The latter can occur when the penalty is sufficiently large. As a result, the lasso, or L1 regression is sometimes used to select the variables to be modelled.

Because L1 regression can shrink coefficients to zero, its use can lead to models that are more sparse than standard regression models and may be easier to interpret as a result. In the proposed investigation, the optimal value of  $\lambda$  is estimated through use of grid search with cross-validation<sup>14</sup>—a process that is handled through use of the R library ‘glmnet’ (Friedman et al., 2021). The resulting optimal penalty term,  $\lambda$ , is applied to all weights except for the intercept.

### The Support Vector Machine (SVM)

As noted above, SVMs have been used with good results by others using acoustic features of speech to infer affect and social signals. In cases where more than two predictors, or features, are used, the SVM learns from the training instances by mapping them to the feature space and then constructing one or more hyperplanes that separate the instances into two classes, forming a decision boundary (Hastie et al., 2009; James et al., 2017).

A hyperplane is a flat affine subspace with one less dimension than the outcome space in which it is embedded such that, assuming a  $p$ -dimensional space, a hyper plane will have  $p-1$  dimensions (James et al., 2017). As a result, in a two-dimensional space such as a cartesian coordinate system with two axes, the associated hyperplane will be a line. In a three-dimensional space, such as a coordinate system with three-axes, the associated hyperplane will be a plane.

Following Hastie et al. (2009), the notion of a decision boundary can be formalized by describing a typical binary classification scenario in which there exists an  $n \times p$  matrix  $X$  comprised of  $n$  observations in  $p$ -dimensional space,

$$\mathbf{x}_1 = \begin{pmatrix} \mathbf{x}_{11} \\ \vdots \\ \mathbf{x}_{1p} \end{pmatrix}, \dots, \mathbf{x}_n = \begin{pmatrix} \mathbf{x}_{n1} \\ \vdots \\ \mathbf{x}_{np} \end{pmatrix}, \quad (2.6)$$

---

<sup>14</sup> Use of the cross-validation framework, or  $k$ -fold cross validation framework, provides a means to test the performance of the study’s models without use of a new sample of data. It is usually employed in what is referred to as the ‘model selection’ phase of the model development process. The  $k$ -fold cross validation process (KCV) consists of splitting the training data into  $k$  independent subsets. Typically, all but one of the resulting  $k$  subsets are used to train the given model and the remaining subset is used to test the resulting model by evaluating the accuracy of the model’s classifications. The training and testing process is repeated until each fold of the data has been used to test the given model and any accompanying hyperparameters. A range of values for any hyperparameters can then be tested without having to make use of the hold out test set.

and a set of  $n$  associated outcomes that fall into two classes such that  $y_1, \dots, y_n \in \{-1, 1\}$  where -1 identifies one class and 1 identifies the second class.

Classification using a hyperplane assumes it is possible to construct a plane with  $p-1$  dimensions such that it separates the training observations according to their respective class labels, in this case -1 and 1. Such a separating hyperplane has the property that on one side of the boundary, the class labels have a value of -1, and on the other side of the boundary, they have a value of 1. Again, following the notation of Hastie et al. (2009), in the case of a two-dimensional outcome space, such a hyperplane has the following properties:

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} > 0 \text{ if } y_i = 1 \quad (2.7)$$

and

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} < 0 \text{ if } y_i = -1 \quad (2.8)$$

Where such a hyperplane is possible, it can be used as the basis for a classifier.

Beyond simply identifying the position of an observation relative to the hyperplane, its perpendicular distance from the hyperplane can also provide information about its label. When the magnitude of the perpendicular distance between an observation and the hyperplane is large, then the observation is located far away from the hyperplane and one can be more confident about its class assignment. Conversely, when the distance between a hyperplane and a given observations is small, confidence in its associated label is less justified.

Once one or more hyperplanes have been constructed, use of the SVM allows previously unexamined instances to be mapped to the feature space and their distance from the existing, learned hyperplane(s) can be evaluated. These new instances can then be labeled depending on their position and distance from the hyperplane(s). The distance from the given instance perpendicular to the given hyperplane can be used to inform the certainty of the resulting classification (James et al., 2017).

In the case of the support vector classifier the resulting hyperplanes are linear (James et al., 2017). A distinguishing feature of the support vector *machines* is that they create a *non-linear* decision boundary using either a radial kernel or a polynomial kernel with a specified degree. A non-linear support vector machine with a radial kernel is employed here.

### **Optimization of the Support Vector Machine**

The support vector machine presents two parameters that must be tuned to maximize its ability to accurately separate classes of observations in a manner that generalizes to new data. These hyperparameters are *cost*,  $c$ , and the hyperparameter  $\gamma$ . When constructing one or more hyperplanes, the location and shape is determined by optimizing against two competing objectives. On the one hand, generalizability of the SVM can be improved where the distance between the hyperplane(s) and the classes of observations is maximized in the training set. On the other hand, accuracy of the model is improved by maximizing the number of observations that are correctly classified in the training set. This trade-off here is the generalizability of the model and its accuracy, which is partially controlled by the value assigned to the *cost* hyperparameter,  $c$ , which adds a penalty for each misclassified data point.

When the value of  $c$  is small, the associated penalty for misclassifications is also small. This results in larger margins between the hyperplane(s) and classes, but also results in a greater number of misclassifications. By contrast, when the value of  $c$  is large, so is the penalty for

misclassification of observations. As a result, there are fewer misclassifications, but the margin(s) are also narrower. At the extreme, overfitting can result with large values of  $c$  and model performance can be expected to decline when run on data other than the training set.

The hyperparameter  $\gamma$  is used with the support vector machine, which specifies non-linear hyperparameters. Informally,  $\gamma$  can be understood to determine the influence of single observations. Large values for  $\gamma$  can result in construction of hyperplanes that are overfit to a small number of observations closely clustered together. On the other hand, values for  $\gamma$  that are very low result in hyperplanes that do not adjust to the complexity of the data and thus risk underfitting.

As carried out here, optimal values for the cost and  $\gamma$  parameters of the SVM are determined through use of a grid search implemented within a cross-validation framework. This allows empirical discovery of values for the two hyperparameters. Development and estimation of the support vector machines was carried out using the Caret package in R (Kuhn, 2008).

### Model Evaluation

An approach is required to evaluate performance of the study's models in accurately classifying each of the 5-second audio clips. Ideally such an approach would be usable even in cases where the data exhibits an imbalance in classes (i.e. one or more classes are more prevalent than another). It should also provide a means for comparing performance of current models against historical efforts by other researchers. Schuller et al. (2012) have advocated for use of two metrics to meet these requirements: unweighted average recall, and the AUC, or the area under the Receiver Operating Characteristic curve (ROC). Unweighted average recall can be used in settings where there is class imbalance, and it is the metric adopted in much of the literature treating detection of affect and social stance from paralinguistic features of speech (Schuller et al., 2012; 2013). Motivation for utilizing the AUC also derives from its extensive use automated detection of both social signals and emotion, allowing for comparison of past and current efforts (Schuller et al., 2012). In the current study, both metrics will be calculated by applying the optimized models to the test set. Unweighted average recall and the AUC are described below.

### Unweighted Average Recall

As given in Equation 8a., a model's recall is defined as the proportion of true positive classifications made by a given model to the sum of its true positive (TP) *and* false negative (FN) classifications:

$$Recall = \frac{TP\ X}{TP\ X + FN\ X} \quad (2.9a)$$

Calculated for competence-focused speech, recall is the total number of audio clips correctly identified (*true positives*) divided by the total number of competence-focused clips correctly identified as such *plus* the number of competence-focused clips inaccurately classified as *likability-focused clips (false negatives)*. Because there are two classes of interest in the current study, recall values can be calculated for competence-focused clips and for likability-focused clips. Those values can then be averaged, giving the unweighted average recall. Stated more formally, given two classes of observations, X and its compliment, unweighted average recall can be specified as,

$$UAR = \frac{1}{2} \left( \frac{TP_X}{TP_X + FN_X} + \frac{TP_{\sim X}}{TP_{\sim X} + FN_{\sim X}} \right) \quad (2.9b)$$

where:

- UAR is the unweighted average recall;
- $TP_X$  is the number of accurate classifications of class X made by the model;
- $FN_X$  is the number of false negative classifications of class X made by the model;
- $TP_{\sim X}$  is the number of accurate classifications of the compliment,  $\sim X$ , made by the model;
- $FN_{\sim X}$  is the number of false negative classifications of the compliment made by the model.

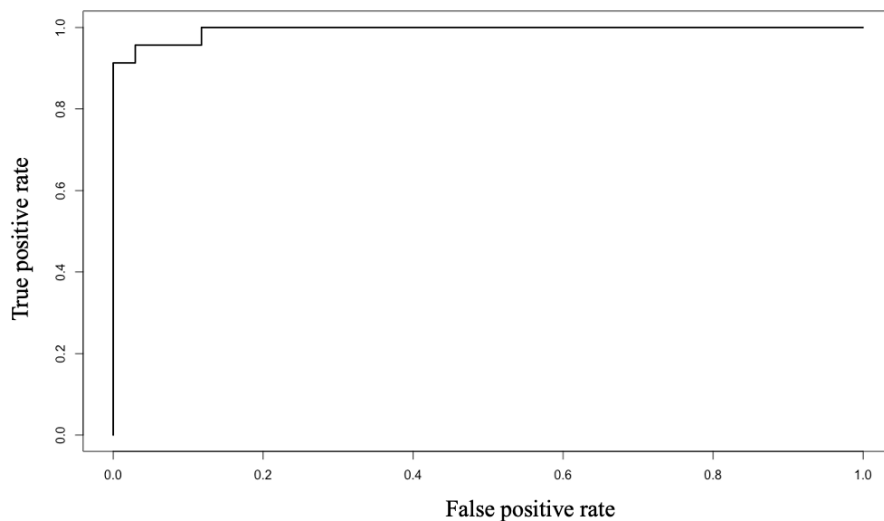
### Area Under the Curve

The AUC, or area under the curve is strictly used for binary classification problems. It is a single value indicating the area under the Receiver Operator Curve (ROC). The ROC is a plot of the true positive rate of a model versus the false positive rate calculated for all threshold values for a model (Hajian-Tilaki, 2013; James et al., 2013). An AUC value of 0.5 indicates a model is performing close to chance. A value of 1 indicates the model is perfectly classifying cases, and a value of 0 indicates it is inverting all classes. A sample receiver operator curve is presented in Figure 2.3, for reference. The ROCR package (Singh et al., 2005) is used to calculate the AUC values for each of the current study's models.

As stated above, benchmark values for unweighted average recall and AUC employed here are inherited from work by Schuller et al. (2012) as part of the INTERSPEECH Challenge for 2012. With the intention of setting a benchmark for the field, Schuller et al. (2012) used a random forest classifier to achieve an unweighted average recall of 0.59 with an AUC of 0.647 in a binary classification task classifying speech as likable or not likable. Their results are close to but better than chance, indicating both the difficulty of the general problem of inferring social signals from acoustic features of speech and the need for continued work in this area.

**Figure 2.3**

*Sample Receiver Operator Curve*



*Note.* The true positive rate is also referred to as ‘sensitivity’. True positives are test or model results that correctly identify the presence of a condition or characteristic. The false positive rate is a test or model results that mistakenly identify the presence of a condition or characteristic when it is not present. The true positive



rate is the proportion of true positives to the total of true positive results *and* the total of all false negative results:  $TPR = TP / TP + FN$ . The false negative rate is the proportion of false negatives to the total number of false negatives *and* true positives:  $FNR = FN / FN + TP$ .

### **Sensitivity and Specificity**

While not a part of the benchmark metrics, values for the sensitivity and specificity of the models are also provided. Sensitivity is the proportion of instances of the primary class that are correctly identified as such by the model, in this case the primary class is competence-focused speech. The sensitivity of a model ranges from 0 to 1. A value of 0 indicates no primary classes were correctly identified, and a value of 1 indicates all primary classes were correctly identified. Specificity indicates the proportion of secondary classes that are correctly identified as such, in this case audio clips made in response to prompts for likability-speech. Specificity takes on the same range of values as sensitivity. Values of 0 and 1, respectively, also indicate none or all instances of the secondary class were detected.

### **Results**

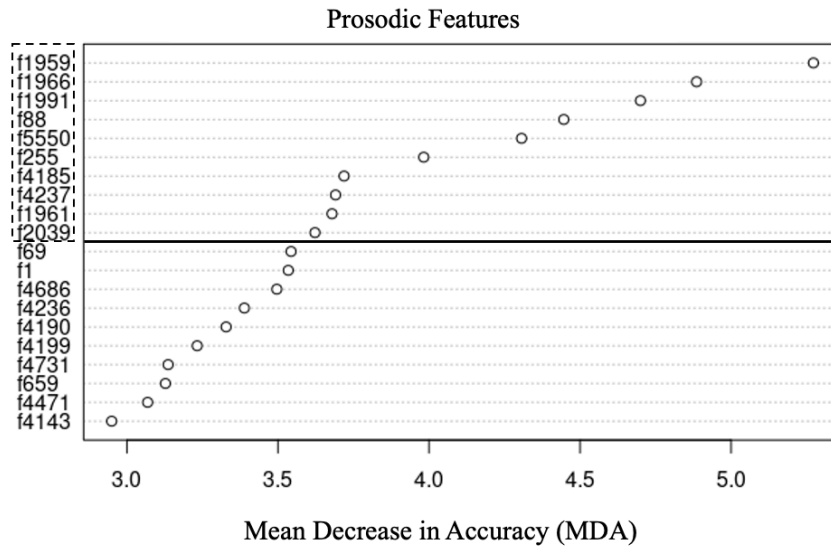
Estimation of feature importance and development of the study's models comprise the analyses carried out in the current chapter. Here, results of those analyses are presented with two emphases. First, the prosodic features that best distinguish competence-focused and likability-focused speech are identified and described. While correlations between the top performing features and the outcome are small, patterns arise regarding differences in the pitch and energy of speech sound in recordings from the competence-focused and likability-focused speaking tasks. It is also notable that in the current study, in which speakers made recordings in their own homes or other every-day environments, a majority of the top performing features utilized a type of filter recognized for isolating speech in potentially noisy environments (Rasta filtering). Second, performances of the study's models are described with reference to their unweighted average recall and AUC values. Performances of the models are related to existing benchmarks from Schuller et al. (2012), as described previously. While model accuracies approach those achieved in Schuller et al., they are still close to chance. Further, the specificity and sensitivity of the models indicate that they are failing to correctly identify many examples of competence-focused speech and are generating false instances of the likability class.

### **Variable Importance**

The top performing features were identified through estimation of the mean decrease in accuracy (MDA) of the random forest model upon removal of the given feature. All features describe an aspect of the prosody of speakers' speech. The mean decrease in accuracy values for the top twenty features are provided graphically in Figure 2.4 with the top ten features identified through use of the dashed box and the horizontal line through the figure indicating the top ten. Table 2.4 subsequently lists the top ten prosodic features in order of their impact on the model accuracy along with the correlations between each of the variables and the outcome, the type of task. The direction of the relationship between each of the top performing acoustic features and the speech label are depicted in the final column.

**Figure 2.4**

*Variable Importance via MDA Estimates of the Top Performing Acoustic Features*



Note: Features in the figure are ordered by increasing MDA values, left to right. The features with the highest MDA values lie above the horizontal line across the figure. Reference names for the acoustic features are provided on the y-axis. For a description of the top ten features, see Table 2.4, first column.

Examining Table 2.4, it is notable that the correlations between the acoustic features and the ground-truth labels range from -0.052 to +0.056, indicating that the associations are weak (Cohen, 2013). Speakers do not appear to be altering any single feature in order to advance competence or likability but rather seem to be adjusting several properties of the speech sound depending on which type of prompt they are responding to. While no single feature of the speech sound can be used to explain how speakers are behaving differently in the competence and likability scenarios, the most effective features relate to differences in either the pitch or the energy of the speech sound. Seven of the top ten features summarize an aspect of the speech sound’s pitch. The remainder summarize an aspect of the speech sound’s energy.

**Table 2.4**  
*Point Biserial Correlations for Features With the Highest MDA Estimates and the Outcome*

Acoustic Feature	Reference	Pt.Biserial	Relation	
audSpec_Rfilt_sma_de[0]_quartile3 numeric Pitch related. Duration audio clip exhibits speech sound in lower three quartiles of the audio spectrum.	f1959	-0.022	Comp-Focused	↓
			Lik-Focused	↑
audSpec_Rfilt_sma_de[0]_stddev numeric Pitch related. Variability in the frequency of the speech sound.	f1966	+0.017	Comp-Focused	↑
			Lik-Focused	↓
audSpec_Rfilt_sma_de[1]_iqr1-2 numeric Pitch related. Duration audio clip exhibits speech sound ranging from the 1 <sup>st</sup> to 2 <sup>nd</sup> quartiles of the audio spectrum.	f1991	-0.051	Comp-Focused	↓
			Lik-Focused	↑
pcm_RMSenergy_sma_lpgain numeric Energy related. Energy of speech signal before processing.	f88	+0.033	Comp-Focused	↑
			Lik-Focused	↓
audSpec_Rfilt_sma_de[0]_posamean numeric Pitch related. Mean of positive values only of the frequency across the audio clip.	f5550	+0.012	Comp-Focused	↑
			Lik-Focused	↓
pcm_RMSenergy_sma_peakMeanAbs numeric Energy related. Mean of peaks in energy across the audio clip.	f255	-0.052	Comp-Focused	↓
			Lik-Focused	↑
audSpec_Rfilt_sma[0]_meanFallingSlope numeric Pitch related. Duration of the audio clip exhibiting falling pitch.	f4185	+0.056	Comp-Focused	↑
			Lik-Focused	↓
audSpec_Rfilt_sma_de[0]_iqr2-3 numeric Pitch related. Duration audio clip exhibits speech sound ranging from the 2 <sup>nd</sup> to 3 <sup>rd</sup> quartiles of the audio spectrum.	f4237	-0.012	Comp-Focused	↓
			Lik-Focused	↑
audSpec_Rfilt_sma_de[2]_risetime numeric Pitch related. Total duration audio clip exhibits rising frequency.	f1961	-0.042	Comp-Focused	↓
			Lik-Focused	↑
pcm_RMSenergy_sma_iqr1-2 numeric Energy related. Duration audio clip exhibits low energy levels (1-2quartiles)	f2039	-0.051	Comp-Focused	↓
			Lik-Focused	↑

*Note.* The audio spectrum represents sound waves in terms of their amplitude (measured in decibels, dB) at a range of frequencies (measured in kHz). Features utilizing measurements of the audio spectrum of speech sounds (audSpec) summarize sound patterns in terms of the amplitude of the sound present at each frequency.

In addition, a majority of the top ten features utilize Rasta filtering (noted as ‘Rfilt’ in the table), a filtering approach that suppresses components of the recordings that change more rapidly or more slowly than human speech sounds (Hermansky & Morgan, 1994). This makes Rasta filtering a popular approach in the field of speech detection, particularly when data is collected in noisy environments. Application of the Rasta filter can be an effective way to remove non-speech sounds from recordings to improve energy estimates, as it filters out acoustic wave forms that change too rapidly or too slowly to have been generated from the human vocal tract, indicating they are a source of noise.

The features that utilize Rasta filtering also relate to speakers’ pitch. Duration of the clips spent in the lower three quartiles of the audio spectrum (f1959) was negatively correlated with competence-focused speech. A similar negative correlation was found between duration of the clip spent in the 2<sup>nd</sup> and 3<sup>rd</sup> quartiles of the frequency spectrum and competence-focused speech (f4237). On the other hand, variability in pitch (f1966) and high average positive frequencies across a clip (f5550) are positively correlated with competence-focused speech.

The remaining three top-performing prosodic features summarize information about the root mean square energy of speech sound in the set of audio clips. Given the symmetric shape of acoustic waveforms above and below the time-axis, the mean values of their amplitude over time will always be zero and are therefore uninformative. The RMS approach provides a way of calculating the average amplitude of the speech sound over time that is informative by calculating the average squared value of the energy and then taking its square root (Kent et al., 2002) so that values other than zero are possible. While two of the three energy-related features indicate that louder speech is correlated with competence-focused speech (f2039 and f88), the fact that having high positive peaks in energy (f255) is negatively associated with competence-focused speech suggests the relation may be more complicated.

### Model Evaluation

Table 2.5 summarizes the model performances on the test set with regard to their unweighted average recall and AUC values. Examination of the table suggests two consistent themes. First, the models' performances are in line with the current benchmark provided by Schuller et al. (2012). Second, while that is the case, it is also true that the models' low sensitivity values indicate they are doing a poor job correctly identifying the positive class, which, in this case, is competency-focused speech.

**Table 2.5**

*Classification Performance Metrics for the Test Set With Non-Expert Speakers*

	Sens	Spec	UAR	AUC
L1-Logistic Regression	0.145	0.924	0.590	0.602
Support Vector Classifier (Linear)	0.018	0.986	0.591	0.548
Support Vector Machine w/ Radial Kernel	0.164	0.877	0.585	0.534

*Note.* Legend: sens: sensitivity; spec: specificity; UAR: unweighted average recall; AUC: area under the curve. The positive class for the models was competence-focused speech. Low sensitivity values and high specificity indicate the models are doing a poor job accurately classifying competence-focused speech.

### L1-Logistic Regression Performance

Performances of the L1-logistic regression as well as the support-vector classifier and support-vector machine are summarized in Table 2.5. The L1-logistic regression was applied to the training data containing the prosodic features with the highest importance values. The modeling procedure utilized a ten-fold cross validation framework and grid search to discover the best performing value of the hyperparameter  $\lambda$ .

The best-performing model set  $\lambda$  at 0.0132. The  $\lambda$  hyperparameter can range in value from 0 to 1. At a value of 0, the L1 logistic regression utilizes all features and returns the least squares fit. At a value of 1, the L1 logistic regression shrinks all coefficients to 0 and returns the null model. At the selected value of 0.0132, the  $\lambda$  parameter resulted in model coefficients of 0.0 for eighty-four of the prosodic features entered into the model.

Given the task of distinguishing competence-focused speech and likability-focused speech, the L1-logistic regression demonstrated an unweighted average recall of 0.633 and an AUC of 0.583 on the training data. Sensitivity of the model, indicating its ability to correctly identify true positives (in this case likability-focused speech) was 0.183. Specificity, indicating the model's ability to correctly identify true negatives (in this case competence-focused speech) was 0.930. This result is noteworthy given the size of the difference between the two values.

When applied to the recordings in the test set, the L1-logistic regression trained classifier had an unweighted average recall of 0.590 and an AUC of 0.583. These values are well aligned

with the benchmark performance from Schuller et al. (2012). Sensitivity for the model with the test set was 0.145 and its specificity was 0.924. This set of values for sensitivity and specificity reflect a similar pattern exhibited with the training set. The majority of the model's classification error stems from a tendency to misidentify likability-focused speech as competence-focused.

### **Support Vector Classifier Performance**

Given the task of distinguishing competence-focused speech and likability-focused speech using the training, the model had an unweighted average recall of 0.574 and an AUC of 0.647. As observed in the case of the L1-logistic regression, the sensitivity for the SVC was very low, 0.001, when using the training data. Its specificity was 1.000. This pattern is continued with the test set. When applied to data from recordings in the test set, the support vector classifier had an unweighted average recall of 0.591 and an AUC of 0.548. Its sensitivity was 0.018 and its specificity was 0.986.

### **Support Vector Machine Performance**

When used with recordings in the training data, the resulting model had an unweighted average recall of 0.704 and an AUC of 0.820. Its sensitivity was 0.361 and its specificity was 0.959. Using the data from recordings in the test set, the support vector machine with radial kernel had an unweighted average recall of 0.585 and an AUC of 0.534. Its sensitivity was 0.164 and its specificity was 0.877. Sensitivity and specificity of the support vector classifier and the support vector machine reflect a similar behavior as that found with the L1-logistic regression: The majority of the models' classification error stems from a tendency to misidentify likability-focused speech as competence-focused.

## **Discussion**

This work represents an effort to automate detection of social signals that indicate whether speech is competence-focused or likability-focused using recordings of extemporaneous speech from non-experts in naturalistic settings. While it is a notably challenging task, investigation of automated detection of social signals in these or similar conditions will be required to make such technologies usable in a broad range of everyday settings. By emphasizing prosodic features of speech, the study also supports investigation of speaker behaviors that differentiate likability-focused and competence-focused recordings. Features relating to energy and pitch of speech sounds consistently differ between recordings made in response to competence-focused prompts and those responding to likability-focused prompts.

Two findings regarding the models' performance are also emphasized here. First, performances of the current study's machine learning models approximate that of the benchmark employed. In order to be utilized in practical settings, efforts to detect social signals such as competence-focused and likability-focused speech, however, will still need to improve. Second, the models developed for the current study consistently exhibit low sensitivities and high specificities, meaning their performance leads to many false negatives of competence-focused speech and many false-positives of likability-focused speech. Additional work should be carried out to create a more balanced model performance with regard to sensitivity and specificity. In what follows, these findings are discussed in more detail, followed by discussion of the study's strengths and limitations.

### **Patterns in Speaker Behavior**

The top performing prosodic features, presented in Table 2.4, exhibited the highest variable importance in terms of their associated mean decrease in accuracy (MDA) values. All ten of those features relate to the measured energy of the recorded speech sound. Those features summarizing aspects of the audio spectrum (audSpec) convey information directly about *both*

energy and frequency, or pitch. Indirectly, they also support inferences about differences in speakers' behavior when responding to the two types of prompts. Review of the top performing features supports three general findings that are summarized in Table 2.6 and described in more detail below. In general, louder speech that varies in its amplitude or 'pitch' is positively associated with competence-focused speech. However, when emphasizing competence-focused speech, variability in amplitude tends to occur through transitions from higher amplitudes to lower amplitudes. Change in the other direction, from low to high amplitude, is associated with likability-focused speech.

**Table 2.6**  
*Patterns of Speaker Behavior and Their Relation to Types of Speech*

Category of acoustic features	Competence-focused speech	Likability-focused speech
Energy	Higher energy	Lower energy
Audio spectrum (energy and amplitude)	Greater variability of speech sound in the audio spectrum	Lower variability in the audio spectrum
Audio spectrum (energy and amplitude)	Longer duration of rising values in the audio spectrum	Longer duration falling values in the audio spectrum

### Energy

The current study found that participating speakers tended to change the energy level of their speech sound when responding to the two types of prompts. Though the relation is weak, louder speech is positively correlated with competence-focused speech. For example, examining feature f88 (the estimated energy of the speech across the clip) speakers tended to speak more loudly when responding to competence-focused prompts and more quietly when responding to likability-focused prompts. This is consistent with data from feature f2039, the duration an audio clip exhibits energy levels in the lower two quartiles. That feature is negatively associated with responses to competence-focused speech and therefore positively associated with responses to likability-focused prompts.

### Audio Spectrum

Second, there is evidence that greater variability of speech sounds along the audio spectrum is positively related to speech made in response to competence-focused prompts and therefore negatively correlated with responses to likability-focused prompts. Higher values for feature f1966, the standard deviation of speech along the audio spectrum (which encompasses measures of both energy and amplitude of speech), indicate variability in the speech sound loudness and/or amplitude and are positively correlated with responses to prompts for competence-focused speech.

Lastly, the direction of change in amplitude of the speech sound seems to be related to the type of speech. Speech with rising values in the audio spectrum is negatively correlated with responses to competence-focused prompts and positively correlated with responses to likability-focused prompts. Evidence for this relation is provided by feature f1961, the duration that a clip exhibits rising frequencies, and its positive correlation with responses to likability-focused prompts. Feature f2039, the duration over which a clip exhibits falling frequency, provides evidence that the converse is also true: speech exhibiting falling frequencies are negatively correlated with likability-focused speech and positively associated with competence-focused speech.

### Patterns in Model Performance

As given in Table 2.5., the AUC values for the current study range from 0.534 to 0.602. Unweighted average recall of the models ranges from 0.585 to 0.591 on the test data. These model performances in the detection of competence-focused and likability-focused responses suggest the models can distinguish the two types of speech at a rate that is better than chance<sup>15</sup>. The results are also in-line with the original baseline performance described earlier in the paper. Schuller et al. (2012) accomplished a mean unweighted average recall of 0.590 and an AUC of 0.647 on the INTERSPEECH test set using a random forest model. The sample of recordings used in that study was a combination of scripted and unscripted speech—primarily discrete statements made in response to automated prompts from a phone-based call-in system, with a mixture of calls made in indoor and outdoor environments (Burkhardt et al., 2010).

While results for the current study are better than chance, and in line with the benchmark performance from Schuller et al. (2012), it is nevertheless notable that the unweighted average recall and AUC values are still quite low. The range of AUC values associated with the models developed for this study are above, but close to chance. This is taken here to indicate the difficulty of the classification task in general and the fact that there is much additional work to do to automate detection of social signals from speech sound.

The fact that the models' performances are so close to chance, also may indicate that there are relationships, or predictors missing from the models. This is to be expected given the study's exclusive focus on prosodic features of speech. Because of its focus on prosodic features of speech, non-prosodic features of the speech sound were intentionally not included here. In subsequent work, these may be leveraged for improved results in a broader, or even brute-force effort utilizing all of the features included in the ComParE dataset for instance.

Further, as mentioned at the outset of the paper, it is expected that individuals simultaneously utilize multiple sources of information to convey and interpret social signals. An appropriate extension of the current study would be inclusion of additional features in the models, inclusion of lexical content of the speakers' recordings, and inclusion of information from modalities other than speech, for example, gesture-related features as well as data representing aspects of facial expression. This approach is taken up in Chapter 3, which presents a set of machine learning models that utilize both acoustic and lexical content of speech.

### Strengths and Limitations

Many of the strengths of the current study originate with its use of non-expert speakers and extemporaneous speech in naturalistic settings. As mentioned in the results section, above, the study's limitations are generally associated with the quality of the machine learning models' performances, the sample of speaker participants, and weak correlations between the study's outcome variables and the prosodic features of speech used as predictors. Each of these points are treated in more detail below.

Only non-expert speakers were selected for the study and all speakers made recordings in their own homes or otherwise familiar and naturalistic environments. These design features are

---

<sup>15</sup> AUC values range from 0 to 1. An AUC of 1 would indicate the given model was identifying both classes of the labels without any error, *i.e.*, no instances of misclassification. An AUC of 0 would indicate that all instances of competence-focused labels were being classified as likability-focused speech, and all instances of likability-focused labels were being misclassified as competence-focused. An AUC of 0.50 would indicate the given model is performing as well as chance.

expected to aid in making the resulting models more generalizable across speakers and social contexts. The study's use of extended extemporaneous responses is also noteworthy. While the prompts used to gather speaker recordings provided clear social situations and explicit directives for the type of speech required, they also allowed speakers to record unique responses. This aspect the study is expected to have yielded greater variability in both the content of the responses and the acoustic features of speech demonstrated by speakers than approaches that utilize scripted speech. It allows for development of models that stand to be more easily generalized to a range of social contexts.

Use of the study's Online-Recording System (ORS), designed by the author to facilitate investigation of social signals, also permits data gathering with relative ease when naturalistic recording environments are desired. The full set of recordings used for the current study required approximately four months to gather with minimal staff time dedicated to the actual collection of the data, though the time requirements associated with data collection did grow with associated quality review of incoming recordings.

The study also presents limitations which should be considered as they impact the accuracy and generalizability of the study's machine learning models. First, while the models developed here exhibit performances that are on par with existing benchmarks, they are not yet performing well enough to be relied on applied settings. It is notable, as well, that the sensitivities of the models are generally quite low. This stands in contrast with the models' specificities, which are much higher. This inverse relation is not unexpected—when sensitivity is low, specificity will naturally be higher, and vice versa. But in practical terms, in cases such as these where sensitivity is very low and specificity is high, the models are doing a poor job of correctly classifying the primary class, in this case competence-focused speech, and are generating a high number of false-positives with regard to the secondary label, likability-focused speech.

As noted in discussion of participants' demographics, while the sample of speakers is somewhat balanced with regard to their reported sex at birth, the majority report they are White and non-Latino. That imbalance is likely to limit the generalizability of the models described here. It is reasonable to assume that variability in patterns of speech sound exhibited by the current sample is lower than it would have been had the sample demographics been more varied.

Lastly, the magnitude of correlations between the candidate prosodic features and the study's class labels are small. The causes for this could be manifold. While other possibilities exist, two candidate explanations should be that features with stronger associations to the class labels were absent in the modeling effort, and that prosodic features alone carry a limited amount of information about speakers' intended social signals. Those possibilities, along with evidence presented in the background of the current paper, suggest there may be benefits to incorporating additional sources of information into the effort to detect social signals.

### **Conclusion**

The current study utilized prosodic features of speech to investigate the feasibility of automating detection of social signals such as competence-focused and likability-focused speech in naturalistic environments with extemporaneous speech from non-expert speakers. It also investigated differences in vocal behaviors of speakers when given tasks that prompted for either competence-focused speech or likability-focused speech. While performances of the models developed here are in line with accepted benchmarks, the models will need to be improved for use in applied settings. Improvements are expected to follow from use of multiple sources of



information other than prosodic features, such as lexical and propositional content of speech, gesture and facial expressions.

## CHAPTER THREE

### Use of Acoustic and Lexical Features of Speech to Detect Social Signals

The preceding chapters focused on detection of competence-focused and likability-focused speech using only continuous acoustic features, or what are also referred to as suprasegmental patterns in the speech sound. The approach did not utilize segmental aspects of speech such as the lexical content of participant recordings, i.e., the words speakers use. In addition, ground truth for the models developed in those same chapters relied on the type of task speakers were responding to. The target inference being made was whether speakers were responding to a task that prompted competence-focused speech or likability-focused speech.

That work comprises an initial step in detecting social signals that are associated with distinguishing competence-focused and likability-focused speech. But it follows an approach that can be improved on at least two fronts. First, using only acoustic features of speech fails to account for important information that is conveyed through its segmental aspects: It ignores *what is said* and accounts only for *the way it is said*. This runs the risk of leaving potentially important information unaccounted for. Second, in many cases, it may be advantageous to develop systems capable of inferring how human observers would likely perceive a speaker. This is not possible when ground truth is defined only in terms of the prompt to which speakers are responding.

A system capable of detecting social signals would ideally account for information derived from both segmental and suprasegmental aspects of speech and would be capable of inferring human impressions. Such a research program will be important for development of robotic systems and virtual agents that partner with or simply respond to humans. One approach to automating inferences about human social signals that may be able to meet these criteria involves development of one or more *inferential detectors*.

Inferential detectors (IDs) provide a means for automated systems to infer the impressions formed by human observers in response to the observers' perceptions of an object, agent, environment, or process (OAPE) (Vallejo et al., 2020). To date, such detectors have been developed to infer human impressions of physical objects and environments such as taste (Jiang et al., 2018), sound (Dal Palu et al., 2014), smell (Staples, 2000), touch (Gee et al., 2005), glossiness (Leloup et al., 2014; Gigilashvili et al., 2019), and color and texture (Eugene, 2008). Inferential detectors have also been created to infer more socially relevant constructs such as individuals' affective states (DeMello & Graesser, 2010; Baker et al., 2012; Fleureau & Guillotel et al., 2012), deceit (Bhaskaran et al., 2011; Ludwig et al., 2016) and flirtatiousness (Ranganath et al., 2009). Through use of a process referred to as fusion,<sup>16</sup> such IDs can be designed to utilize information from multiple sources.

#### Purpose

In what follows, a shift in perspective is made from previous chapters: from inferring speakers' intention to convey one or more social signals, to emphasizing instead interlocutors' likely impressions of speakers. The purpose of the work is three-fold. First the work aims to present a generalized process for development of IDs that infer social signals. Second, the work aims to investigate the feasibility of developing inferential detectors that utilize extemporaneous speech in naturalistic environments and multiple sources of information to infer impressions of social signals that human observers would be likely to form. Third, clips with ratings in either the upper or lower quartiles of raters' scores are used to investigate the features of speech and

---

<sup>16</sup> Fusion includes a set of processes and approaches for joining data from multiple sources, typically sources that span modalities, for use as predictors for one or more predictive or classification models.

therefore the speaker behaviors that best differentiate the two extreme classes. Two inferential detectors are created: one that infers human judgements of audio clips exhibiting low or high levels of competence-focused speech, and another that infers human judgements of audio clips exhibiting low or high levels of likability-focused speech.

Throughout the chapter, emphasis is placed on inferential detection, the processes for developing IDs and demonstration of that process in the context of detecting social signals from speech. Because the chapter investigates features of speech that differentiate competence-focused and likability-focused speech, it will be necessary to introduce some of the basic concepts of acoustic speech analysis. The chapter will not seek provide an introduction to acoustic analysis in general, but rather the chapter will seek to describe enough of the basic concepts of acoustic analysis of speech that the reader will be able to gain an understanding of how the properties of speech sound may be used to detect social signals. Acoustic analysis of speech is a vast and technical topic. It is expected that the brief introduction of basic concepts of acoustic analysis that is offered here is permissible because of the chapter's focus on processes and investigation of inferential detection more broadly.

### **Intended Contributions**

Existing approaches to detection of social signals tend to utilize raw scores of one or more raters as ground truth. The general process for developing IDs that is proposed here makes use of an additional modeling step that maps ordinal scores from human raters onto a continuous scale utilizing a faceted rating scale Rasch model and creating a continuous measure, or interval level scale. Mapping the human ratings onto a continuous scale allows for confirmation and investigation of the distances, or measured differences, between the ordinal values awarded by raters. Resulting estimates can then be grouped to create different levels of measurement, dichotomous or ordinal level variables. Selecting two cut points within the scale to identify the audio clips in the first and fourth quartiles, as is done in the current study, leads to a dichotomous variable, categorizing audio clips as either 'low' or 'high' on the given scale.

Importantly, use of the faceted rating scale model also allows one to account for differences in rater severity and to detect erratic rater behavior. Because the effort utilizes both acoustic features of speech as well as segmental features of speech, the project also affords a comparison of model results when only a single source of information is used (e.g., the lexical content of speech vs. its acoustic features) and when multiple sources of information are used (lexical content *and* acoustic sources).

### **Background**

The background discussion for the current chapter is divided into three general parts. In the first part, some of the basic properties of speech sound are introduced with particular attention given to the intensity and frequency of the sound waves that comprise speech sound. In the second part of the background section, the composite nature of speech is introduced – namely, the fact that speech can be described in terms of its sounds as well as the words that are conveyed. Lastly, the concept of inferential detection is introduced, and a general process is described for development of inferential detectors (IDs). Throughout the background discussion and the chapter as a whole, the sound associated with speakers' vocalizations is referred to as 'speech sound'.<sup>17</sup>

---

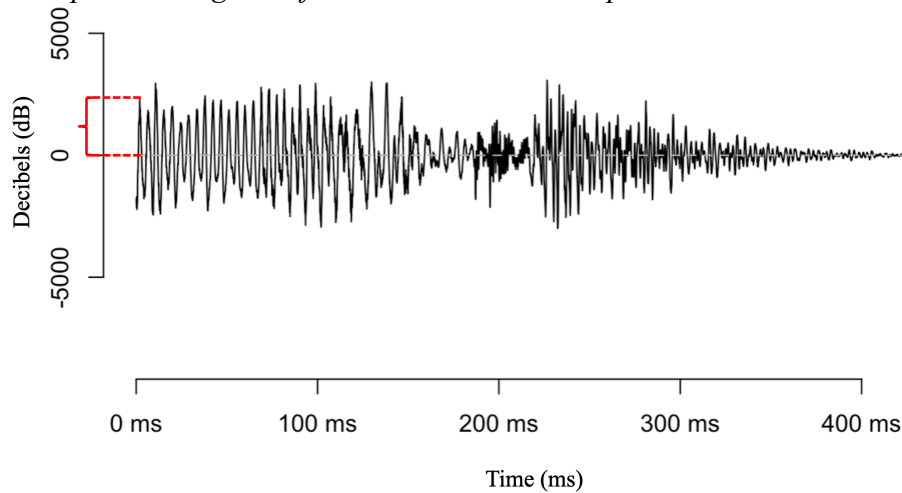
<sup>17</sup> The term, 'speech sound' is both a phenomenon and a subject of study and is therefore treated as a non-count noun. As a result, references will be made to 'speech sound' without use of an indefinite article such as '*a* speech sound'. Use of a definite article may be made such as '*the* speech sound' when referring to a particular instance of speech sound.

When it is considered as a physical phenomenon, speech sound can be identified with a complex airborne wave that is audible by human ears or sensible by a microphone. Speech sound is a complex wave because it is comprised of multiple waves that vary in their amplitude and frequency. The amplitude and frequency of soundwaves making up the speech sound play a role in determining what the speech sound sounds like to human observers.

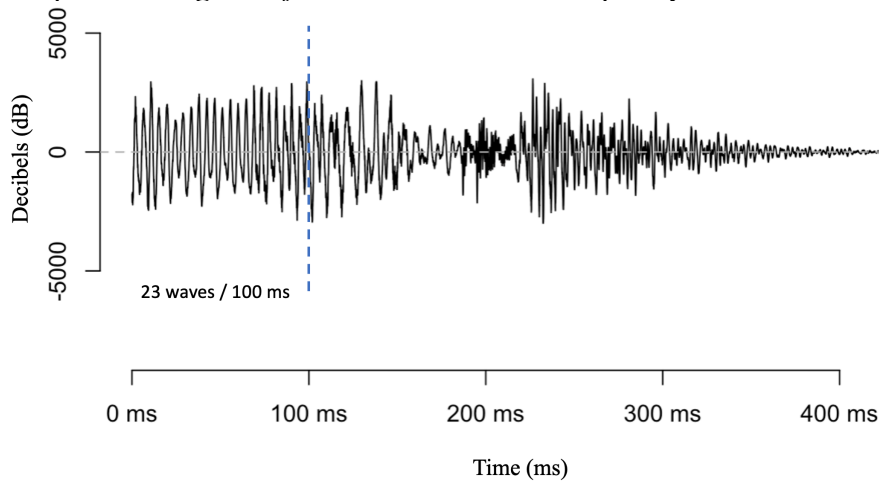
The amplitude of soundwaves making up speech sound is often measured in decibels and is perceived as loudness. The amplitude of soundwaves from a sample of speech sound can also be visualized through use of an oscillogram. Oscillograms visualize speech sound as a transverse wave form that has a changing amplitude, measured in decibels, that varies over time, usually measured in milliseconds. An example of an oscillogram is provide in Figure 3.1a which highlights the amplitude of the first wave with a red bracket. The oscillogram shows the first wave has an amplitude of approximately 2,500 decibels. This particular oscillogram was generated from an excerpt of a digital recording of a speaker reading the phrase “Lend me a nickel.” The oscillogram visualizes the amplitude of the soundwaves for the word ‘nickel’.

**Figure 3.1**

*Figure 3.1a. Sample Oscillogram of the Word ‘Nickel’, Amplitude Noted*



*Figure 3.1b. Sample Oscillogram of the Word ‘Nickel’, Frequency Noted*



The frequency of soundwaves making up speech sound is measured in Herz (Hz). A soundwave that has a frequency of 680,000 hertz is vibrating at a rate of 680,000 times per second. In Figure 3.1b, the oscillogram representing speech sound from a recording of the word ‘nickel’ has a frequency of 23 vibrations in the first 100 milliseconds, or 230 Hz. Soundwaves with higher frequencies are perceived as having a higher pitch. Soundwaves with lower frequencies are perceived as being lower in pitch.

Speakers can manipulate properties of their speech sound to convey meaning. As an example, speakers may change the amplitude, or loudness of their speech sound, speaking quietly when making a polite request and speaking very loudly when making an angry demand. Speakers may also alter the frequency of their speech sound in order to speak with a higher or lower pitch. Alternatively, speakers can also coordinate simultaneous changes both amplitude and frequency of their speech sound as in the case of someone expressing surprise or confusion by starting with a quiet low frequency vocalization that rises in both amplitude and frequency across the word, “Whaaaaat?!!”

### **The Source-Filter Theory and Formant Frequencies**

As stated above, speech sound is comprised of multiple sound waves that have multiple frequencies. The source-filter theory of speech production provides an explanation for this fact by describing the production and modification of speech sound in terms of the anatomical components of the human vocal tract. Initially, air is pushed from the lungs and passed over the vibrating vocal folds of the larynx (the *source*) to create sound waves. The sound waves generated by the vibrating vocal folds are then modified by different parts of the vocal tract (*filters*). Components of the vocal tract include the larynx, epiglottis, tongue, lips, the oral cavity, and the nasal cavity, among others. Modulation of soundwaves generated at the vocal folds occurs as the result of coordinated motor activity involving these components of the vocal tract (Fant, 1960; Kent & Read, 1992).

The rate at which the vocal folds of the larynx vibrate determines what is called the *fundamental frequency*, or  $F_0$ , of the resulting speech sound. The fundamental frequency is informally associated with the pitch of a person’s voice. As a person speaks, or sings for that matter, a person can alter the pitch of their voice. Without necessarily being aware of it, speakers are altering the rate at which their vocal folds are vibrating when they change the pitch of their voice. When the vocal folds vibrate quickly, the resulting sound waves have a higher frequency and are therefore perceived to have a higher pitch. When the folds vibrate more slowly, the resulting soundwaves have a lower frequency, and are therefore perceived to have a lower pitch.

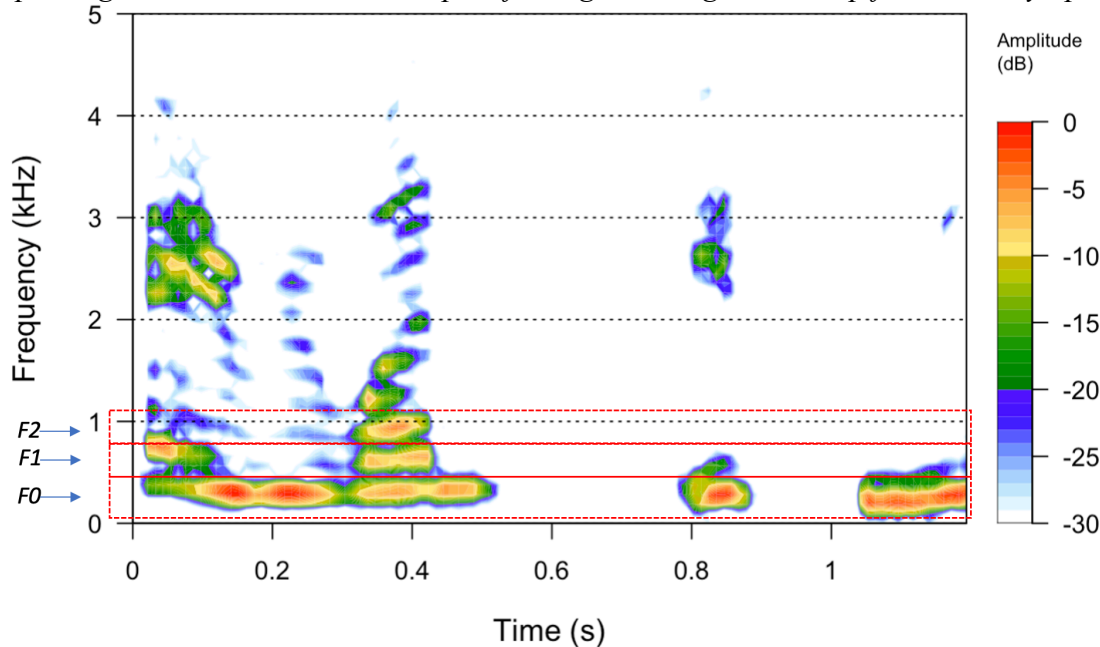
As the sound waves generated by the vibrating vocal folds pass through the vocal tract they can begin to resonate in different ways. As the sound waves resonate along the vocal tract, the sound waves’ frequencies are altered, leading to sound waves that have frequencies that are different from the original, fundamental frequency. The frequencies of such altered sound waves are referred to as *formant frequencies* ( $F_1, F_2 \dots F_n$ ). The fundamental frequency is always the lowest frequency. The  $F_1$  formant frequency is the second lowest, the  $F_2$  formant frequency is the third lowest frequency, etc. Thus, in a spectrogram which visualizes the various frequencies of a sample of speech sound and how those frequencies change over time, the  $F_0, F_1$  and  $F_2$  frequencies appear to be layered, or ‘stacked’, one above the other, starting with the fundamental frequency at the bottom. This order is demonstrated in Figure 7 which presents a spectrogram of a 5 second clip of recorded speech in which the speaker is trying to emphasize their likability.

The amplitude of the fundamental frequency as well as the amplitude of the formants is encoded in Figure 3.2 through use of color. Red areas exhibit the highest amplitude, or loudness.

Blue areas in the figure exhibit the lowest amplitude. At approximately 0.2 seconds into the recording for example, one can see that soundwaves at the fundamental frequency have the greatest amplitude, i.e., are loudest. At approximately 0.4 seconds however, the amplitude of the fundamental frequency is similar to that of the sound waves at the first and second formants, *F1* and *F2*. In general, the soundwaves associated with the first formant, *F1*, are initiated through resonance occurring between the larynx and the back of the tongue. The second formant, *F2*, is generated by resonance occurring within the oral cavity. Both *F1* and *F2* are associated with generation of sounds relevant for voicing vowels.

**Figure 3.2**

*Spectrogram Created From a Sample of a High Scoring Audio Clip for Likability Speech*



*Note.* The spectrogram visualizes patterns of speech sound that arise from its frequency, amplitude, and temporal aspects. The fundamental (*F0*) and formant frequencies (*F1* and *F2*) are indicated with red boxes. In theory, there can be an infinite number of formant frequencies. Typically, only the first and second are emphasized. In this sample, the fundamental frequency is a concentration of energy in the speech sound occurring at approximately 0 to 0.3 kHz. The first formant, *F1*, ranges from approximately 0.3 to 0.8 kHz, with the second formant stacked on top of that, at approximately 0.8 to 1.2 kHz.<sup>18</sup>

Speakers can modify the amplitude and frequency of their speech sound within milliseconds, exhibiting remarkably fine-grained control. Just as it is possible to exhibit such fine-grained control over the amplitude and frequency of one's speech sound, it is also possible to measure changes in properties of a speaker's speech sound that occur within milliseconds. The overall loudness or amplitude of the speech sound as well as its pitch, or fundamental frequency can each be measured. Likewise, the amplitude of the speech sound at specific frequencies can

<sup>18</sup> Examining Figure 7, the reader is asked to notice that the frequency and the amplitude of the speech sound can and do vary over time. Thus, rate of change of those quantities can also be a feature of speech, as well as the range of those values over a given unit of time. While the focus here is on the acoustic features of speech, it is helpful to remember that the speech sound is the result of complex motor activity across multiple physical systems that humans begin exploring and learning to manipulate at early ages. Competence-focused and likability-focused speech can therefore be explored in terms of their associated motor activity, forming a conceptual link between social signals and embodied activity.

be determined as well. Because such summaries of the speech sound can be carried out at specific intervals of time, it is also possible to summarize speech sound in terms of the range and variation it exhibits in both amplitude and frequency over time. Measurement of many other properties of a speaker's speech sound is possible too. Working with the assumption that speakers vary these and other properties of their speech sound in order to convey one or more social signals, it is expected that measurements of such properties can aid in detecting those same social signals.

### **Double Articulation**

While the sound waves comprising speech sound can be thought of as being continuous, speakers utilize them to form discrete words. This characteristic allows speech sounds to be characterized in terms of two sets of simultaneously occurring sound patterns—a phenomenon referred to as “double articulation” (Hockett & Hockett, 1960; Filippi, 2016). The first set of sound patterns arise through changes in amplitude and frequency of a speaker's speech sound (Lehiste, 1976). Patterns arising from changes in amplitude and frequency of speech are referred to as suprasegmental patterns of speech. Suprasegmental patterns of speech comprise the so-called song, or prosody, of speech. The second set of patterns arise through concatenation of individual speech sound segments, *phonemes*. By themselves, phonemes are meaningless. But they can be grouped to form word units and words that *do* have meaning (i.e., *morphemes* and *lexemes*, respectively) and they can be used to express propositional content (Hauser et al. 2001).

Speakers, regardless of culture and language, modulate both aspects of the speech sound—its segmental and its prosodic, or *suprasegmental*, aspects (Wildgruber et al., 2006; Ackerman et al., 2016; Filippi, 2016). Together, they comprise *what* is said and *how* it is said, allowing speakers to convey messages about a range of topics such as external states of affairs, their emotions and attitudes, as well as their readiness and ability to coordinate with others (Filippi, 2016). In what follows, segmental patterns of speech are identified with the words that speakers use, and they are understood to comprise the *lexical* content of speech. Suprasegmental patterns of speech are identified with the broadly *acoustic* content of speech. With the assumption that both sets of patterns can provide nonredundant information, work to detect social signals may benefit by accounting for the fact that speakers make use of both the acoustic and lexical aspects of speech to communicate.

### **Use of Acoustic and Lexical Features in Speech to Convey Social Signals**

The idea that both acoustic and lexical features of speech contain information about social signals of speakers is consistent with research into speech comprehension more generally. That work provides evidence that comprehension of speech involves integration of multiple sources of information. Among others, these sources include the lexical and related propositional content of speech, its prosodic content, as well as visual information that includes both facial expressions as well as gesture or body movement in general (Massaro & Cohen, 1983; Massaro, 1987; Wildgruber et al., 2006; Ackerman et al., 2014).

Further, if efficient solutions are preferred in nature (Sutherland, 2005), the fact that multiple sources of information are used in speech comprehension would suggest that the different sources may not be completely redundant and/or they may perform an amplification function. Lastly, practical successes among recent efforts to infer speaker affect, attitudes, and social signals indicates use of multiple sources of information leads to improved outcomes for such efforts (Vinciarelli et al., 2009; Wagner et al., 2011; D'Mello & Kory, 2015).

Given these points, a system capable of detecting social signals would ideally account for multiple sources of information. In the case of inferring social signals such as those involved in

portraying one's competence or one's likability, it may be beneficial to account for both segmental and suprasegmental aspects of speech. Such a research program will be important for development of robotic systems and virtual agents that partner with or simply respond to humans.

One approach to automating inferences about human's social signals that may be able to meet these criteria involves development of one or more *inferential detectors*. Inferential detectors can be created that utilize information from multiple sources – such as suprasegmental and segmental features of speech. In the following section, inferential detectors are introduced and a general process for their creation is described.

### **Inferential Detectors**

Inferential detectors, also referred to as virtual sensors, utilize a set of measurement processes that allow for quantification of properties usually determined by human perception or judgement (see Vallejo et al., 2019). Their development generally requires three scopes of work: 1) selection and measurement of physical properties of a target object, agent, process, or environment (OAPE); 2) measurement of human perceptions or judgements of those properties; and 3) development of one or more models capable of creating a mapping between the two resulting sets of measured values. When all three scopes of work are carried out successfully, the resulting inferential detector(s) can be used to justify claims about how the target OAPE *would be* perceived or judged by a human observer, given the target's physical properties, even in the absence of such an observer. Such claims are justified with reference to the quality of the two sets of measurements involved and the quality of the mapping function(s) between them.

There exist several examples of such inferential detectors (IDs). They have been developed to infer human impressions of objects and environments such as taste (Jiang et al., 2018), sound (Dal Palu et al., 2014), smell (Staples, 2000), touch (Gee et al., 2005), glossiness (Leloup et al., 2014; Gigilashvili et al., 2019), and color and texture (Eugene, 2008). Inferential detectors have also been created to infer perceptions and judgements of others and their behavior, such as individuals' affective states (D'Mello & Graesser, 2010; Baker et al., 2012; Fleureau & Guillotel et al., 2012), instances of deceit (Bhaskaran et al., 2011; Ludwig et al., 2016) and flirtatiousness (Ranganath et al., 2009). In each case, their development relies on the same general processes.

### **A General Development Process for Inferential Detectors**

As described in the previous section, the general process for developing inferential detectors can be organized into three categories of effort: 1) Gathering instances of the target OAPE and measuring its relevant physical properties; 2) deriving measured values that indicate the intensity of human perceptions or judgements of the target; 3) determining an appropriate mapping between the measured physical properties of the target object or agent and the measurement of the intensity of human impressions. While they are presented here as distinct phases of work, it should be noted that the first two phases can be treated in parallel and that the full scope of work can be conducted iteratively, allowing teams to put in place detectors that are sufficient for their desired use and that can be improved over time. A generalized description of the scopes of work associated with development of an inferential detector for inferring human impressions is described here and summarized in Figure 3.3a, below. A more specific summary of that process for detection of social signals from speech is provided in Figure 3.3b.

#### ***Phase 1***

In the first phase, the emphasis is placed on identifying or collecting instances of the target OAPE. In the case of inferential detectors for social signals such as those entailed by

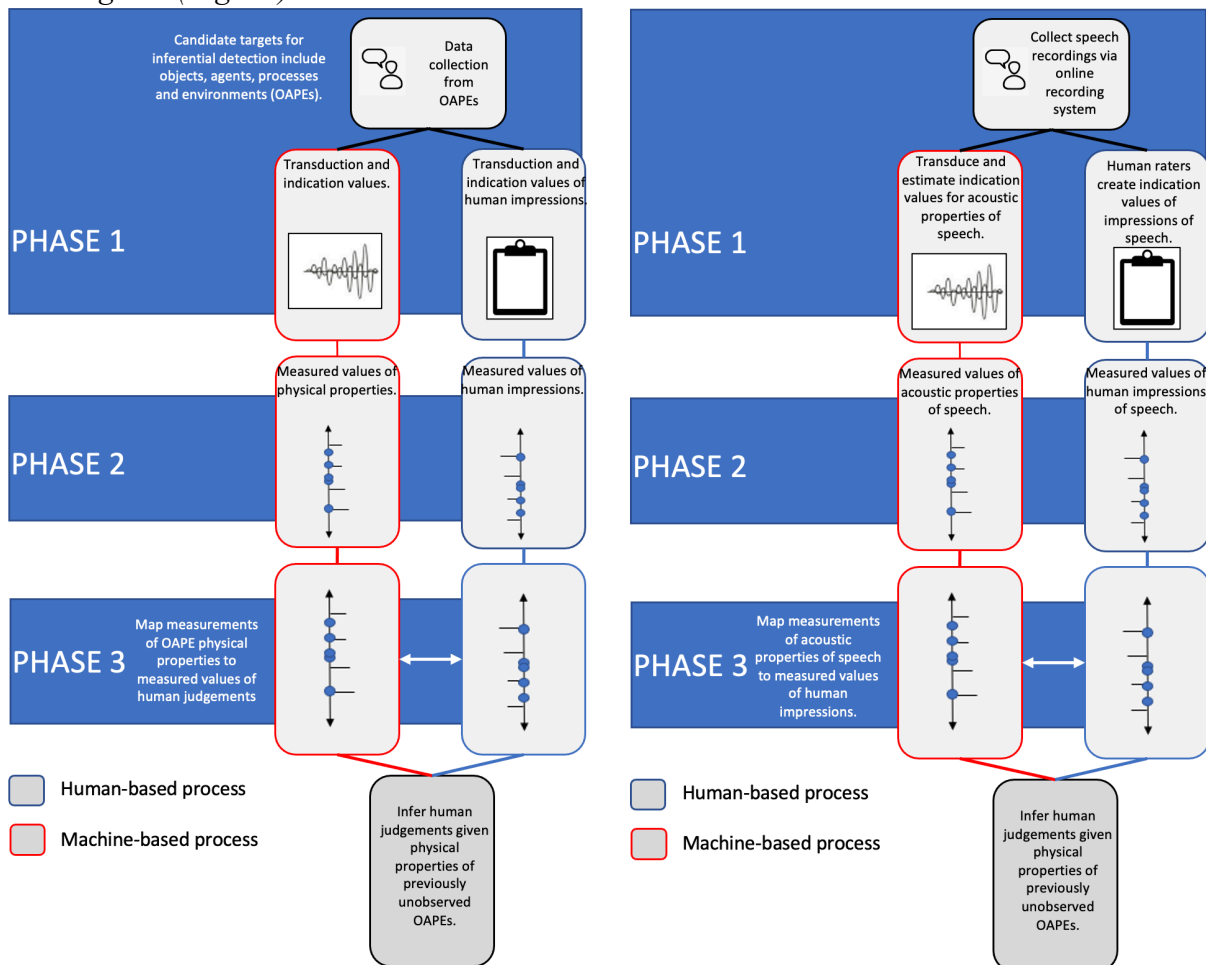


competence-focused and likability-focused speech, collection of such instances might result in a corpus of audio recordings or audio-video recordings from individuals exhibiting the desired behaviors. Phase 1 also includes aspects of the measurement process with a series of transductions occurring that transform the physical properties of the target into indication values. On the machine-based side of the process, one or more instruments can be used that transform physical properties of the target into quantitative values by indicating the extent to which the property brings about physical changes to the transducer. In the case of audio recordings of participant speech, one of these transducers could be a microphone, for instance. Using current technology, transduction of the acoustic properties of recordings can also occur through digital means.

On the human-based side of the process, human observers utilize one or more ordinal or dichotomous rating instruments to assign a numeric value that associates a magnitude with their impression(s) of the target. These ratings can also be thought of as indication values. The perspective held here is that the human observers transduce the acoustic signals conveyed in speech recordings, for example, as raters observe and then form an impression of those recordings. The raters' act of making a mark on the rating scale(s) assigns a quantitative rating to their experience. As is typically the case, such indication values are not in a form that meets requirements of measurement (Mari et al., 2021). When observers use ordinal scales for their indication values, for example, the order of the values may increase or decrease across the scale, but not monotonically, so that ordinal values making up the scale(s) are not necessarily equidistant. What is required is an approach to calibrating the indication values that are generated at this phase of the ID process to one or more continuous scales that meet requirements for measurement.

**Figure 3.3**

*General Development Process for Inferential Detectors (Fig. 7a) as Applied to Detection of Social Signals (Fig. 7b)*



*Note.* Figure 3.3a. General development process for inferential detection.

*Note.* Figure 3.3b. Development process for inferential detection as applied for inferential detection of social signals from speech—competence-focused and likability-focused speech.

**Phase 2**

In Phase 2, the indication values generated in Phase 1 are mapped onto a continuous interval scale using one or more measurement models. In the case of physical properties of sound recordings (such as those used in the current study, scales such as energy, measured in decibels, for example, or pitch measured in amplitude) the scales and units used may be familiar. In these cases, the task is to map physical changes in the transducer(s) (e.g., vibrations per second) to the relevant scale to create one or more measured values.

In the case of human ratings, the mapping task is similar. There is an additional complication, however, as it cannot be expected that human raters behave mechanistically, transducing the acoustic signals in the exact same manner as other human raters. Raters will exhibit variability

amongst themselves in the way they interpret or apply the rating scales. Some raters will require more of the property to be present than others before awarding a high indication value while others will require less of the property to be present to award the same high value. This variability can be thought of as rater severity; severe raters consistently require more of the property to be present than other, less severe raters, before awarding an equivalent indication value, or score. Similarly, some raters will be inconsistent in the way they transduce the speech recordings, resulting in inconsistent or even erratic rating behavior. An approach is required that can identify and account for both types of unexpected rating behavior. In the process presented here, the faceted rating scale model (FRSM) (Linacre, 1989) is used to identify and account for these sources of variability. The faceted rating scale model and its use in the ID process is discussed in more detail in the analysis section of this chapter.

### ***Phase 3***

In Phase 3, models are developed that map the measured values of physical properties (that were estimated in Phase 2) to measured values indicating the presence or intensity of human perceptions or judgements. The end point of this phase of the ID process is development of one or more models that are capable of ingesting new observations of physical properties of an agent expressing a social stance and mapping those to a set of quantified values and/or classifications that accurately reflect human impressions. In such a case, the mapping process is aligned with conventional supervised learning processes in which the human ratings can serve as ground truth, and quantified values of the physical properties associated with participating agents' actions serve as candidate model features.

### **Inferential detectors for Competence-Focused and Likability-Focused Speech**

In this chapter, this three-phase process is followed to develop inferential detectors capable of predicting human judgements of audio clips that exhibit either high or low levels of competence-focused or likability-focused speech. The approach can make use of information from acoustic as well as lexical features of speech. Three sets of machine learning models are developed. The first utilizes acoustic features of speech only; the second, lexical features of speech only; and the third utilizes a combination of both acoustic and lexical features. These three models are also compared.

### **Materials and Methods**

Requirements of transparency and openness of the work described here are met by providing access to the source code for the analyses, given in Appendix C.1.

### **Study Design**

This study uses a retrospective cross-sectional design to investigate automated detection of likability-focused and competence-focused speech from expert and non-expert speakers who responded extemporaneously to prompts in naturalistic, uncontrolled environments. A stratified random sample of five-second audio-clips was drawn from the larger pool of participants' recordings and subsequently delivered for review and scoring by human raters. After raters' scores were modelled, the audio-clips judged to exhibit the highest (4<sup>th</sup> quartile) and those judged to exhibit the lowest (1<sup>st</sup> Quartile) levels of the two speech types were selected and used for development of a set of machine-learned models. The resulting models utilize acoustic and/or lexical features of the audio-clips to infer the human raters' judgements.

### **Participants**

The stratified random selection of five-second audio clips used for this study were drawn from recordings by the expert (n = 101) and non-expert (n = 154) speakers described in previous chapters. Expert speakers were recruited using online markets Fiverr (<http://fiverr.com>) and

Upwork (<http://Upwork.com>). The bios for the speakers were reviewed for evidence of training and/or professional experience in acting.

As described in Chapter 2, non-expert speakers were recruited through the Amazon Mechanical Turk system (<http://mturk.com>). The MTurk setup option was used to limit participation in the study to those workers who had achieved the formal status of high-performing workers, further recognized as masters within the MTurk system (Peer et al., 2014). Additionally, a geographic restriction was implemented that permitted only individuals confirmed as current residents of the United States to participate in the study. Candidate nonexpert speaker participants responded to a demographic survey prior to being accepted to the study, and only those indicating that they were native English speakers of American English were allowed to participate.

### **Speech Recordings**

Speaker participants were provided access to the study's Online Recording System (ORS). This cloud-based recording tool presented speaking prompts embedded in an online graphic novel. The ORS requests access to the microphone on the user's computer and leads them through a series of checks to ensure the microphone is working correctly and that the noise level of the user environment is acceptable. Users then complete a series of recording prompts that are embedded in an online graphic novel titled *Advice Hour*, in which participants assume the role of a podcast host responsible for coaching callers on how to respond to a series of communication challenges in their work and private lives. Each scenario in the graphic novel prompts the speaker-participants to record what the caller should say in the scenario, in the manner they should say it, exhibiting competence-focused or likability-focused speech. A sample of the recording prompts is included in Appendix C.2. After each recording, the ORS gives speakers a chance to review their recording and either accept or revise it. Recordings were automatically saved in a secure cloud-based file.

### **Speech Pre-Processing**

Participants' recordings were reviewed by members of the research team for evidence of on-task performances. Recordings for each task were then segmented into five-second clips and indexed by a unique identifier indicating the speaker, speaker-type (expert versus non-expert), task, task type, and the window rank of the clip (a cardinal value denoting the position of the clip within the full recording).

### **Audio-Clip Selection**

Audio clips were selected for review and scoring by raters through a two-stage stratified random sampling process. In the first stage, a stratified random sample of sixty ( $n = 60$ ) speaker participants was drawn from the combined pool of speaker-participants. Reported sex at birth and level of expertise (expert versus nonexpert) informed the strata in this first draw. In the second stage, a random draw of four hundred five-second clips was made from recordings by this group of selected speakers. One or more raters identified forty-six audio clips that had poor audio quality and were subsequently dropped from the sample. The final pool of three hundred and fifty-four ( $n = 354$ ) audio clips was reviewed and rated by all eight raters.

### **Audio Clip Review and Rating**

Eight members from the study team trained to rate the study's selected audio clip. The eight-member team comprised four females and four males. Training followed a three-stage process. In the first stage, study team members assembled a corpus of publicly available audio recordings that they felt exemplified either competence-focused or likability-focused speech. The team then jointly reviewed the clips, identifying, discussing, and documenting those perceived

acoustic features of the clips that distinguished instances of competence-focused speech from instances of likability-focused speech. For competence-focused speech, the team identified perceived features of speech that the group identified with efforts to emphasize one’s intelligence, motivation, and energy. For likability-focused speech, the team identified perceived features of speech they associated with friendliness, warmth, and care. The rating team then independently scored the pilot clips on each of the study’s rating scales, described in Table 3.1<sup>19</sup>. Team members then discussed and adjudicated their scores in cases where there were disagreements. The team followed this process through five iterations, using new audio clips from the pilot each time, until achieving a minimal targeted level of agreement of 0.70 for each of the rating scales, using Cohen’s weighted kappa (Cohen, 1968; Banerjee, 2008).

Once rater training was completed, the eight members of the study team reviewed and rated the pool of four-hundred five-second clips selected for the study. As presented in Table 3.1, raters recorded their impressions of each audio clip by responding to a series of eight rating scales—four for competence-focused speech and four for likability-focused speech. Each scale presented raters with four possible levels of response ranging from 0 (*Not at all*) to 3 (*Very much*). Ratings were based on raters’ individuals’ impressions of the speech presented in the given audio clip.

**Table 3.1**  
*Rating Scales Quantifying Raters’ Impressions of Speech*

<i>The audio-clip sounds like the speaker is emphasizing their . . .</i>							
intelligence	motivation	energy	overall competence	friendliness	warmth	care	overall likability
3.Very much	3.Very much	3.Very much	3.Very much	3.Very much	3.Very much	3.Very much	3.Very much
2.Moderately	2.Moderately	2.Moderately	2.Moderately	2.Moderately	2.Moderately	2.Moderately	2.Moderately
1.Slightly	1.Slightly	1.Slightly	1.Slightly	1.Slightly	1.Slightly	1.Slightly	1.Slightly
0.Not at all	0.Not at all	0.Not at all	0.Not at all	0.Not at all	0.Not at all	0.Not at all	0.Not at all
<b>Rating scales for competence-focused impressions</b>				<b>Rating Scales for likability-focused impressions</b>			

**Data Structure and Data Processing**

Two sets of files were created and used for the eventual analyses presented here: one for estimation of measured values from raters’ scores, using the faceted rating scale model (FRSM), and a second for development and testing of the planned machine learning models for classification of speech types—competence-focused and likability-focused speech.

***File format for Estimates Using the FRSM***

Two data files were created from raters scores: one for estimates of competence-focused speech, and the second for estimates of likability-focused speech. In both cases, data was organized into a long file format<sup>20</sup> as required by the TAM package for R (Robitzsch et al.,

<sup>19</sup> As discussed in the analysis section of the current paper, use of the term ‘rating scale’ here is referring to the scales used by raters to assign quantitative values, or indication values to their impressions of the speech sound in a given recording. The term is also used in discussion of the faceted rating scale model.

<sup>20</sup> As demonstrated in Tables 2a. and 2b., ‘long’ file formats allow for repeated presentation of data associated with a single participant or single observation within the data file. The long file format is commonly used when multiple observations are made of the same participant over time, or when multiple raters judge a single performance. Both cases apply here. Multiple five-second audio clips are gathered from individual speakers. In turn, audio-clips may be judged by multiple raters. The ‘long’ file format can be contrasted with the ‘wide’ format displayed in Tables 3a. and 3b. In that case, data for each audio clip is presented along a single row.

2022), used to estimate raters' severity and the extent to which audio-clips were perceived to have emphasized the target speech type. Each five-second audio clip was associated with rating values from the eight raters. Ratings were indexed by rater-id, audio clip-id, and speaker-id. Samples of the file format for competence-focused speech and likability-focused-speech are presented in Tables 3.2a and 3.2b.

### Table 3.2

#### *File Formats for Ratings Data*

Table 3.2a. File Format for Estimating Rater Harshness and Scores for Competence-Focused Speech

Clip-id	Rater-id	Intelligence	Motivation	Energy	Overall-Competence
1357	R1	1	2	1	1
1357	R2	1	2	2	2
1357	R3	3	2	1	2
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

Table 3.2b. File Format for Estimating Rater Harshness and Scores for Likability-Focused Speech

Clip-id	Rater-id	Friendliness	Warmth	Care	Overall-Likability
1357	R1	1	2	1	1
1357	R2	1	2	2	2
1357	R3	3	2	1	2
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

### File format for the Classification Task

Raters' scores were used to estimate the extent to which audio clips exhibited competence-focused or likability-focused speech. A faceted rating scale model (described in detail in the discussion of the study's models) was used for this purpose. Audio clips falling in the upper and lower quartiles of these estimates were selected and used to develop a set of machine-learned models capable of classifying audio-clips based on whether they exhibit either high levels or low levels of the targeted speech type.

Files for the classification task were organized in wide format, as is typically required by the R packages used in the current study. Each row of the file included data for a single audio clip, its outcome label, associated lexical features, and associated acoustic features, all indexed by the audio clip-id and the speaker-id<sup>21</sup>. A sample of the file format is presented in Tables 3.3a and 3.3b.

<sup>21</sup> As noted in the previous chapter, data in this and similar studies is necessarily nested with 5-second audio clips nested within longer recordings which are in turn nested within recording tasks and speakers. The fact that the data is structured in this manner is not treated in the modeling stage. Ideally, that structure would be incorporated into the models used. Unfortunately, few multi-level versions of popular machine-learning models are currently available in existing libraries, though this is beginning to change in applications of machine learning in education (Cannistra et al., 2021) and public health (Ji et al., 2020), e.g.

**Table 3.3***File Formats for Speech Classification*

Table 3.3a. File Format for Model Development for Classifying High and Low Levels of Competence-Focused Speech

Speaker-id	Clip-id	Competence-focused label	Mutual information value for positive class (High-level)	Mutual information value for negative class (Low-level)	Acoustic feature $f_1$	...	Acoustic feature $f_n$
001	13579	high	0.746	-1.4	0.216		-4.500
001	14021	low	1.689	0.988	0.153		1.830
002	18560	low	-0.868	0.921	1.342	.	-2.500
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
132	97555	low	-3.572	1.654	0.0532		0.7721

Table 3.3b. File Format for Model Development for Classifying High and Low Levels of Likability-Focused Speech

Speaker-id	Clip-id	Likability-focused label	Mutual information value for positive class (High-level)	Mutual information value for negative class (Low-level)	Acoustic feature $f_1$	...	Acoustic feature $f_n$
001	26842	low	0.009	-2.1	0.112		-4.500
001	32549	high	2.768	1.389	0.886		0.003
002	12437	low	-0.444	-0.878	2.112	.	-5.198
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
132	124.37	low	-0.868	0.921	1.346		-2.500

### Outcome Definition

Measured values of the targeted speech type are estimated using the faceted rating scale model described in the modeling approach section of the current chapter. For detection of high and low levels of competence-focused speech, a single binary outcome label was created from the resulting measured values in order to indicate whether a given five-second audio clip was in the fourth quartile of measurements (high-level) or in the first quartile of measurements (low-level) for the given speech type. The positive class indicates clips in the highest scoring group, the fourth quartile, and the negative class indicates clips in the lowest scoring group, the first quartile. The same approach was used for detection of high and low levels of likability-focused speech.

### Analysis

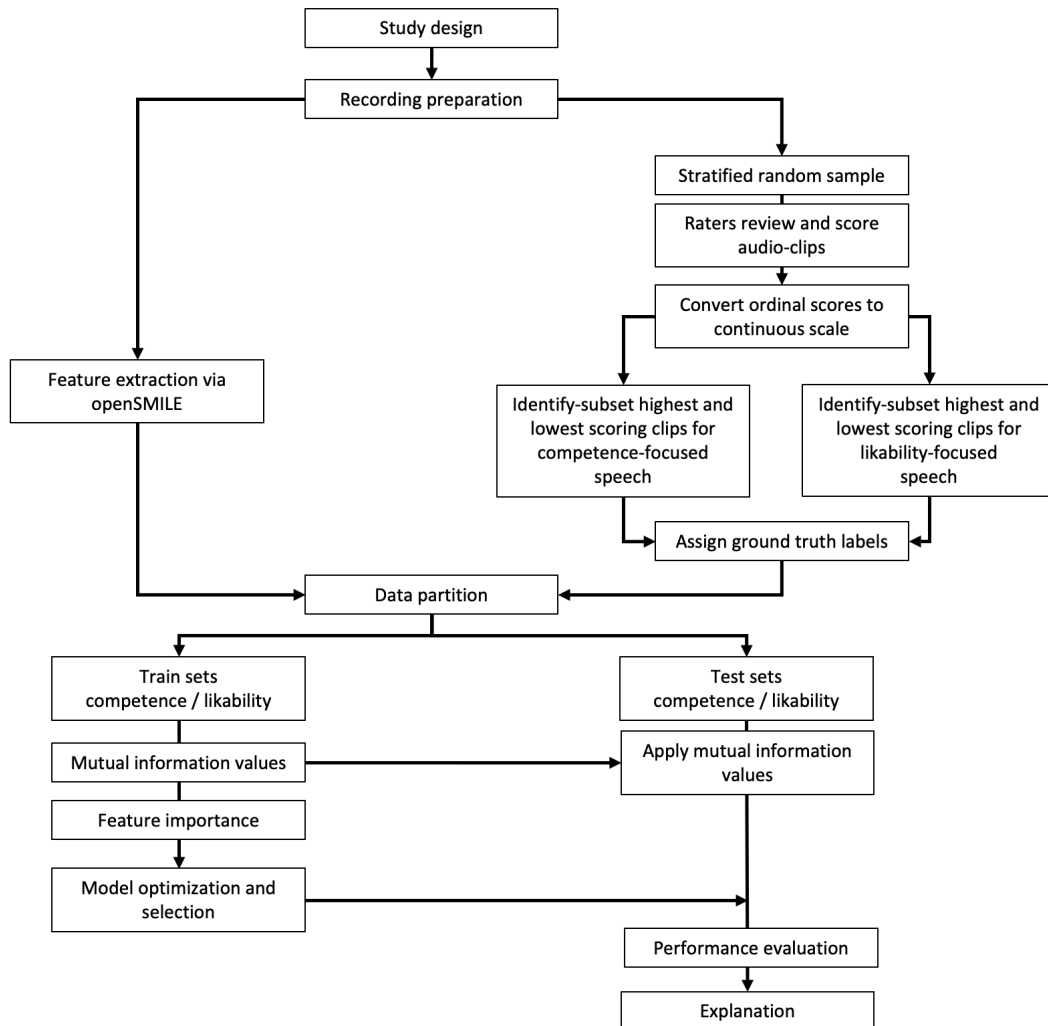
#### Data Preparation

A schematic of the data handling and analysis pipeline is presented in Figure 3.4. After preparation of the digital audio recording files and extraction of the acoustic features included in the Computation Paralinguistics Evaluation feature set (ComParE) via the Python-based openSMILE toolkit,<sup>22</sup> each resulting acoustic parameter was joined with its respective

<sup>22</sup> openSMILE (Eyben et al., 2013) is a Python based toolkit for extraction of acoustic features of speech. SMILE is an acronym for *Speech and Multimedia Interpretation by Large-space Extraction*. The toolkit facilitates extraction

taskID, window rank, recordingID and unique speakerID as part of the data preparation process. In addition, as a part of the data preparation process, the file was checked for missing values, and none were found. All acoustic parameters were standardized resulting in a mean of 0 and a standard deviation of 1.

**Figure 3.4**  
*Data Handling and Modeling Pipeline for the Study*



### Sampling, Rating, and Labelling Procedures

A total of 4,713 audio clips were gathered as a part of a larger study. Subsequently, a stratified random sample of  $n = 400$  audio clips was drawn from the full set for review and a final set of  $n = 356$  audio clips was used for rating and analysis. The sex of speakers at birth and their expertise level (non-experts versus actors) were used as strata to ensure balance across these two demographic characteristics. Members of the study team rated the resulting three hundred

---

of a wide range of direct and indirect measures of the acoustic properties of speech. The ComParE feature set is one set of features that can be extracted using the openSMILE toolkit. It is comprised of 6,373 direct and indirect measures of properties of speech sound.



and fifty-six clips, indicating the extent to which they thought speakers emphasized competence-focused and likability-focused speech. The rater scores were then converted to an interval level variable for both constructs through use of the faceted rating scale model, as discussed in the analysis section of the current chapter.

Using the resulting interval level values, clips with scores in the upper or lower quartiles for competence-focused speech and those in the upper or lower quartiles for likability speech were identified and labelled. This resulted in two data files for further analyses: one for audio-clips exhibiting either the highest or the lowest levels of competence-focused speech, and a second for audio-clips exhibiting either the highest or the lowest levels of likability-focused speech.

### Table 3.4

#### *Demographics of Selected-Speaker Participants*

Table 3.4a. Reported Demographics for Speaker-Participants in the Train and Test Sets for *Competence-Focused Speech*

	Competence-Focused Speech			
	Train		Test	
	N	%	N	%
<b>Total</b>	37	100	16	100
<b>Sex</b>				
Female	20	54.1	7	43.8
Male	17	45.9	9	56.2
<b>Expertise</b>				
Expert	16	43.2	10	62.5
Non-expert	21	56.8	6	37.5

Table 3.4b. Reported Demographics for Speaker-Participants in the Train and Test Sets for *Likability-Focused Speech*

	Likability-Focused Speech			
	Train		Test	
	N	%	N	%
<b>Total</b>	39	100	16	100
<b>Sex</b>				
Female	19	48.7	9	56.3
Male	20	51.3	7	43.7
<b>Expertise</b>				
Expert	21	53.8	8	50.0
Non-expert	18	46.2	8	50.0

### Data Partitions

Data was partitioned using a 70:30 train-test split, with random selections made at the speaker level to avoid leakage of information between the resulting train and test data sets. Audio clips from a total of 53 speaker participants make up the file for competence-focused speech analyses, and audio-clips from 55 speakers make up the file for likability-focused speech analyses. Details for the speaker participants are provided in Tables 3.3a and 3.3b.

### *Competence-Focused Speech Data*

The training set for competence-focused speech represents a total of 112 audio clips; 51 (46%) clips are from the fourth quartile of scores; the remaining 65 (54%) of the audio clips are

from the first quartile of scores.<sup>23</sup> The testing set represents a total of 56 audio clips, with 31 (55%) from the fourth quartile of scores and 25 (45%) from the first quartile of scores. Labels are *high* and *low*.

### **Likability-Focused Speech**

The training set for likability-focused speech represents a total of 116 audio clips, 51 (44%) labeled high-level and 65 (56%) labeled low-level. The testing set represents a total of 50 audio clips, 33 (66%) labeled high-level and 17 (34%) labeled low-level. Samples are well balanced with regard to sex at birth but exhibit less balance with regard to speaker status—i.e., expert vs non-expert speakers.

## **Modeling Approach**

### **The Faceted Rating Scale Model (FRSM)**

The faceted rating scale model from the Rasch item response theory framework was employed to convert raters' ordinal level scores to continuous values on a linear scale before identifying the upper and lower quartiles for classification. The faceted rating scale model applies the many facet Rasch model developed by Linacre (1989), with the rating scale model (Andrich, 1978) for ordinal level scores. Both models are consistent with the Rasch modeling framework. A benefit of the many facets Rasch approach is that the resulting linear scales are invariant to the rating scales used and the specific raters making the judgements (see Wright & Masters, 1982 for additional detail on this point).

Following Eckes (2011), when applying the faceted rating scale model, the probability of a rater assigning a score, or indication value, to an audio clip is given by:

$$\log \left[ \frac{p_{nijk}}{p_{nijk-1}} \right] = \theta_n - \delta_i - \alpha_j - \tau_k \quad (3.1)$$

where:

- $p_{nijk}$  is the probability of audio clip  $n$  being rated with an indication value  $k$  by rater  $j$  on the rating scale item  $i$ ,
- $p_{nijk-1}$  is the probability of audio clip  $n$  being rated with an indication value  $k-1$  by rater  $j$  on the rating scale item  $i$ ,
- $\theta_n$  is the estimated measured value indicating the extent to which the speech sounds recorded in clip  $n$  exhibit the targeted social stance (in this case competence-focused or likability-focused speech),
- $\delta_i$  is the estimated measured value indicating the difficulty of rating scale (item)  $i$ ,
- $\alpha_j$  is the estimated severity of rater  $j$ ,
- $\tau_k$  is the difficulty of receiving rating  $k$  relative to rating  $k-1$ .

The,  $\tau_k$  term can be identified with the *threshold* between two adjacent scores, and it is the point at which two adjacent scores are equally probable. In the current analysis, each of the

---

<sup>23</sup> Data in this and similar studies is clustered with 5-second audio clips nested within longer recordings which are in turn nested within recording tasks and speakers. The fact that the data is structured in this manner is not treated in the modeling stage. Ideally, that structure would be incorporated into the models used. Unfortunately, few multi-level versions of popular machine-learning models are currently available in existing libraries, though this is beginning to change in applications of machine learning in education (Cannistra et al., 2021) and public health (Ji et al., 2020).

four rating scales for competence-focused and those for likability-focused speech are the same. More specifically, the rating scales all utilize three thresholds to create four ordered groups of scores that carry the same interpretation: 0 v 1, 1 v 2, and 2 v 3. Use of the rating scale model as in the current analysis reflects this situation, and the category coefficients,  $\tau_k$ , are calibrated jointly across the four rating scales for competence-focused speech and the four rating scales for likability-focused speech.

#### Notes on the Use and Interpretations of the FRSM in the Current Study

Special attention should be given to the interpretation of the  $\theta_n$  and  $\delta_i$  terms in the faceted rating scale model in this instance. Since their conception, the original Rasch model, and the many facets Rasch model have been used extensively in the context of assessment of individuals' knowledge, skills and abilities. In the context of such assessments, the  $\theta_n$  and  $\delta_i$  terms have been traditionally interpreted as estimates of a given test taker's ability and the difficulty of a given task, respectively. However, other meanings can be attributed to the two parameters, depending on the context and manner in which the models are applied.

In the current effort,  $\theta_n$  is an estimate of the extent to which the speech sounds presented in a given 5-second audio clip exhibit the targeted social signal, either competence-focused or likability-focused speech. Said differently, in this study  $\theta_n$  refers to a property of the speech sound of a given audio clip. That property is the extent to which the given speech sound of an audio clip exhibits the targeted speech type, or social signal. The speech sound is the *object of measurement* and the extent to which the speech exhibits likability-focus or a competence-focus is the *measurand*, or the property being measured.

Thus, a value for theta is estimated for each 5 second audio clip in the study. In cases where the faceted rating scale model fits the data, high values for theta indicate audio clips that exhibit high levels of the targeted speech type. Low values for theta indicate audio clips with low levels of the targeted speech type. As a result, where  $\theta$  is sometimes referred to as the 'person' parameter because it is interpreted as an estimate of person ability, here it may be thought of as the 'speech sound' parameter as it is an estimate of the extent to which the speech sound of a given audio clip exhibits the targeted social signal – either competence-focused or likability-focused speech.

On the other hand, the  $\delta_i$  term in the current study is an estimate of the difficulty of the given rating scale. Formally it is parallel to the usual idea of a test or survey item. There is one  $\delta_i$  estimate for each rating scale used by the study's raters. As described in Table 3.1, a total of eight rating scales were used by the study's human raters. Four of the scales were used to assign quantitative values to raters' impressions of the extent to which the speech sound in a given audio clip exhibits *competence*-focused speech. The remaining four scales were used to assign quantitative values to the raters' impressions of the extent to which the speech sound in a given audio clip exhibits *likability*-focused speech. It may be more difficult for raters to assign a value of three on one scale for example, than it is for the same raters to assign the same value on a different scale. It is important to identify and account for such differences in the difficulties of the rating scales in order to accurately estimate the extent to which a given clip exhibits the targeted social signal.

Further, special attention should also be given to use of the term 'rating scale' in the current work. The term 'rating scale' is used here to refer to the scales used by raters to assign a quantitative value, or indication value, to their impression of the speech sound associated with a given audio clip. The eight rating scales used by raters are described in Table 3.1. As mentioned previously, each rating scale is a four level Likert scale. The term 'rating scale' is also used in

description of the faceted rating scale model – a type of Rasch model used here to estimate the extent to which the study’s audio clips exhibit the targeted speech type.

Another important aspect of Linacre’s faceted approach to the rating scale model is that it does not require different raters to agree to the same indication values, or rating scale scores in response to a given audio clip. Instead, estimation of the faceted rating scale model requires consistency *within* raters so that the resulting estimate of their severity,  $\alpha_j$ , serves as a summary measure of their rating style (Linacre, 1989; Eckes, 2011). This reflects the conviction that individual raters are experts at forming impressions of social signals *and* that they may differ in their impressions of the same samples of speech.

The TAM package for R (Robitzsch et al., 2022) was used to estimate the parameters for faceted rating scale model. Two runs of the model were made, with one using raters’ indication values for competence-focused speech and the second using raters’ indication values for likability-focused speech. The quartiles for the resulting  $\theta_n$  estimates were identified. Those audio clips exhibiting the highest and those exhibiting the lowest levels of competence-focused speech (the fourth and first quartiles respectively) were labeled as such and used with the study’s machine learning models. The same analysis was carried out with raters’ indication values for likability-focused speech. This allowed binary labels to be associated with audio clips in the upper and lower quartiles for competence-focused ratings and a second set of labels for audio-clips in the upper and lower quartiles for likability-focused speech.

### Feature Importance

The importance of the acoustic features of the selected audio clips was inferred via estimates of the mean decrease in accuracy (MDA) resulting from removal of each feature from a random forest trained on the training data. Random forests are ensembles of classification, regression, or survival trees (Breiman, 2001). Mean decrease in accuracy values for each acoustic parameter were estimated using the Caret package in R (Kuhn, 2008). The approach used is summarized again in this section for convenience.

The process of calculating MDA values using a random forest utilizes permutation of out of bag (OOB) samples to compute the importance of a given variable. OOB samples are observations that were not used in construction of a given tree within a random forest. The collection of OOB observations is used to estimate the prediction error for a given tree and then to evaluate the importance of one or more variables by removing them from the feature set and recalculating the prediction error of the tree (Janitza et al., 2016; Han et al., 2016). For each tree in a random forest, the prediction error (error rate in the case of classification problems) is calculated using the OOB observations. The same calculation is repeated after permuting each feature, or predictor. The differences between the two classification errors—before and after permutation—are averaged over all the trees (Han et al., 2016). Following Janitza et al. (2016) and Han et al. (2016), the equation can be specified as follows:

$$MDA_i = \frac{1}{ntree} \sum_{t=1}^{ntree} (EP_{ti} - E_{ti}) \quad (3.2)$$

where:

- $ntree$  indicates the number of trees in the given random forest;
- $E_{ti}$  indicates the OOB error on tree  $t$  before permuting values of feature  $X_i$ ;
- $EP_{ti}$  indicates the OOB error on tree  $t$  after permuting values of feature  $X_i$ .

This same procedure is repeated for all variables across all trees. Larger MDA values for a given variable indicate its importance for prediction accuracy relative to the other variables used in the random forest model.

### Mutual Information

Extraction of lexical features from the study's audio clips follows the approach of Lee and Narayanan (2002). The general strategy is to identify keywords in each clip that indicate competence-focused and/or likability-focused speech. This is accomplished by estimating the *mutual information* provided by each word in a given audio clip and the clip's label. Words that provide more information about the clip's label (upper quartile or lower quartile of score) appear more often in speech with that category label than in the corpus as a whole (Lee & Narayanan, 2002). The approach is closely aligned with the notions of *self-mutual* information (Cover & Thomas, 2006)<sup>24</sup> and *informativity* (Priva, 2015).

Following Lee and Narayanan (2002), in order to calculate the salience of a word, words in a given clip are denoted by  $W = \{w_1, w_2, \dots, w_n\}$  and the set of speech classes by  $S = \{s_1, s_2, \dots, s_k\}$  (in the present case,  $k = 2$ , negative, i.e., not present, and positive, i.e., present). The mutual information for a single word is given by:

$$i(w_n, s_k) = \log_2 \frac{P(s_k|w_n)}{P(s_k)} \quad (3.3)$$

where  $P(s_k|w_n)$  is the posterior probability that an audio clip containing word  $w_n$  implies speech type  $s_k$ , and  $P(s_k)$  denotes the prior probability of that speech type. Importantly, if a given word  $w_n$  is present in an audio clip and is positively correlated to a speech type label, then  $P(s_k|w_n) > P(s_k)$ , and  $i(w_n, s_k)$  is positive. By contrast, if word  $w_n$  is negatively correlated with speech type  $s_k$ , then  $i(w_n, s_k)$  will be negative. If there is no correlation between the word and the given type of speech,  $i(w_n, s_k)$  will be zero, as  $P(s_k|w_n) = P(s_k)$ . The mutual information of all the words in a given audio clip is given by adding together the mutual information of each word. The notion of mutual information for lexical features of speech can be extended to the mutual information of bigrams, trigrams, and n-grams in general. This extension is not treated in the current work.

### Supervised Machine Learned Models

The L1-logistic regression, support vector classifier, and support vector machine are investigated as candidate approaches to create the desired mapping between the human rater impressions of the study's audio clips and their acoustic and lexical features. Details of these three modeling approaches are presented in Chapter 1 and Chapter 2. Descriptions of the models are provided here for convenience.

#### *The L1 Logistic Regression Classifier*

L1 logistic regression is used to model the probability of a given audio clip being assigned to a competence-focused or a likability-focused prompt label. The model yields a number between 0 and 1 representing the probability of class membership. In the proposed use, the threshold probability (the probability at which an audio clip has an equal probability of either being a member of the given speech type class or not) is set to 0.5.

Assuming the speech type outcome is denoted as  $Y$ , which has a binary outcome that is 0 if the label is not the targeted speech type and 1 if it is, and assuming the predictors, or features, are denoted as  $X$ , the aim is to model the conditional probability that the outcome  $Y$  has a value

---

<sup>24</sup> Self-mutual information, or simply, mutual information, is the information one event provides about another.

of 1 given the predictors  $X$ . This conditional probability is denoted by  $p(Y=1|X)$ . The full logistic regression model can be presented as a regression of the log-odds, so that:

$$\log\left(\frac{p(Y=1|X)}{1-p(Y=1|X)}\right) = \beta_0 + \beta_1 X + \dots + \beta_n X \quad (3.4)$$

where the expression  $\log\left(\frac{p(Y=1|X)}{1-p(Y=1|X)}\right)$  is the logarithm of the odds,  $\beta_0$  is the intercept, and  $\beta_1 \dots \beta_n$  describe the weights associated with each of the modeled predictors (or features) of the given audio clip.

In the supervised machine learning context, the objective is to estimate values of  $\beta_0$  and each of the weights  $\beta_1 \dots \beta_n$ , the sum of which results in a probability of  $X$  that most accurately classifies all the observed data (Hastie et al., 2009; James et al., 2017). Those observations where  $Y$  belongs to the targeted speech type should have a probability as close as possible to 1, and those that do not should have a probability as close as possible to 0.

Following Hastie et al. (2009), this objective can be rephrased in terms of maximizing the product of these two probabilities, i.e., the likelihood:

$$\log\left(\prod_{i:Y_i=1} p(X_i) \prod_{j:Y_j=0} (1 - p(X_j))\right) \quad (3.5)$$

where  $\Pi$  denotes the products over  $i$  and  $j$  which run over the observations classified as 1 and 0 respectively.

Alternatively, one can also rewrite Equation 4 in the form of the *negative* log likelihood:

$$L = -\log\left(\prod_{i:Y_i=1} p(X_i) \prod_{j:Y_j=0} (1 - p(X_j))\right) \quad (3.6)$$

in which case the objective is to estimate the intercept,  $\beta_0$ , and the given weights  $\beta_1 \dots \beta_n$ , by minimizing  $L$ .

### **Optimization of the L1 Logistic Regression Classifier**

L1 logistic regression, or lasso regularization, adds a penalty term,  $\lambda$ , to the log likelihood function:

$$L + \lambda \sum |\beta_1 \dots \beta_n| \quad (3.7)$$

Terms  $\beta_1 \dots \beta_n$  represent features, or measured properties from 1 to  $n$ , and their associated regression weights,  $\beta$ . The term  $\lambda$  is a free parameter, or hyperparameter, with a value that is selected to minimize the error that results when running the eventual model on data comprising the test set, i.e., the out-of-sample error. The lasso accomplishes this by shrinking some of the estimated coefficients, or regression weights, toward or equal to zero. The latter can occur when the penalty is sufficiently large. As a result, the lasso, or L1 regression is sometimes used to select the variables to be modelled.

Because L1 regression can shrink coefficients to zero, its use can lead to models that are more sparse than standard regression models and may be easier to interpret as a result. In the proposed investigation, the optimal value of  $\lambda$  is estimated through use of grid search with cross-validation, a process that is handled through use of the R library `glmnet` (Friedman et al., 2021). The resulting optimal penalty term,  $\lambda$ , is applied to all weights except for the intercept.

### ***The Support Vector Machine (SVM) and Support Vector Classifier (SVC)***

As noted above, SVMs have been used with good results by others using acoustic features of speech to infer affect and social signals. In cases where more than two predictors, or features, are used, the SVM learns from the training instances by mapping them to the feature space and then constructing one or more hyperplanes that separate the instances into two classes, forming a decision boundary (Hastie et al., 2009; James et al., 2017).

A hyperplane is a flat affine subspace with one less dimension than the outcome space in which it is embedded so that—assuming a  $p$ -dimensional space—a hyper plane will have  $p-1$  dimensions (James et al., 2017). As a result, in a two-dimensional space such as a cartesian coordinate system with two axes, the associated hyperplane will be a line. In a three-dimensional space, such as a coordinate system with three-axes, the associated hyperplane will be a plane.

Following Hastie et al. (2009), the notion of a decision boundary can be formalized by describing a typical binary classification scenario in which there exists an  $n \times p$  matrix  $X$  comprised of  $n$  observations in  $p$ -dimensional space,

$$\mathbf{x}_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{1p} \end{pmatrix}, \dots, \mathbf{x}_n = \begin{pmatrix} x_{n1} \\ \vdots \\ x_{np} \end{pmatrix}, \quad (3.8)$$

and a set of  $n$  associated outcomes that fall into two classes so that  $y_1, \dots, y_n \in \{-1, 1\}$  where -1 identifies one class and 1 identifies the second class.

Classification using a hyperplane assumes it is possible to construct a plane with  $p-1$  dimensions such that it separates the training observations according to their respective class labels, in this case -1 and 1. Such a separating hyperplane has the property that on one side of the boundary the class labels have a value of -1, and on the other side of the boundary they have a value of 1. Again, following the notation of Hastie et al. (2009), in the case of a 2-dimensional outcome space, such a hyperplane has the following properties:

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} > 0 \text{ if } y_i = 1 \quad (3.9)$$

and

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} < 0 \text{ if } y_i = -1 \quad (3.10)$$

Where such a hyperplane is possible, it can be used as the basis for a classifier.

Beyond simply identifying the position of an observation relative to the hyperplane, its perpendicular distance from the hyperplane can also provide information about its label. When the magnitude of the perpendicular distance between an observation and the hyperplane is large, then the observation is located far away from the hyperplane and one can be more confident about its class assignment. Conversely, when the distance between a hyperplane and a given observations is small, confidence in its associated label is less justified.

Once one or more hyperplanes have been constructed, use of the SVM allows previously unexamined instances to be mapped to the feature space, and their distance from the existing, learned hyperplane(s) can be evaluated. These new instances can then be labeled depending on their position and distance from the hyperplane(s). The distance from the given instance

perpendicular to the given hyperplane can be used to inform the certainty of the resulting classification (James et al., 2017).

In the case of the support vector *classifier*, the resulting hyperplanes are linear (James et al., 2017). A distinguishing feature of the support vector *machines* is that they create a *non-linear* decision boundary using either a radial kernel or a polynomial kernel with a specified degree. A non-linear support vector machine with a radial kernel is employed here.

### ***Optimization of the Support Vector Machine (SVM) and Support Vector Classifier (SVC)***

The support vector machine presents two parameters that must be tuned to maximize its ability to accurately separate classes of observations in a manner that generalizes to new data. These hyperparameters are cost ( $c$ ) and the hyperparameter  $\gamma$ . When constructing one or more hyperplanes, their location and shape is determined by optimizing against two competing objectives. On the one hand, generalizability of the SVM can be improved where the distance between the hyperplane(s) and the classes of observations is maximized in the training set. On the other hand, accuracy of the model is improved by maximizing the number of observations that are correctly classified in the training set. The trade-off is generalizability of the model and its accuracy, which are partially controlled by the value assigned to the *cost* hyperparameter,  $c$ , which adds a penalty for each misclassified data point.

When the value of  $c$  is small, the associated penalty for misclassifications is also small. This results in larger margins between the hyperplane(s) and classes but also results in a greater number of misclassifications. By contrast, when the value of  $c$  is large, so is the penalty for misclassification of observations. As a result, there are fewer misclassifications, but the margin(s) is also narrower. At the extreme, overfitting can result in large values of  $c$  and model performance can be expected to decline when run on data other than the training set.

The hyperparameter  $\gamma$  is used with the support vector machine, which specifies non-linear hyperparameters. Informally,  $\gamma$  can be understood to determine the influence of single observations. Large values for  $\gamma$  can result in construction of hyperplanes that are overfit to a small number of observations closely clustered together. On the other hand, values for  $\gamma$  that are very low result in hyperplanes that do not adjust to the complexity of the data and risk underfitting.

As carried out here, optimal values for the cost and  $\gamma$  parameters of the SVM are determined through use of a grid search implemented within a cross-validation framework. This allows empirical discovery of values for the two hyperparameters. Development and estimation of the support vector machines was carried out using the Caret package in R (Kuhn, 2008).

### **Performance Metrics**

An approach is required to evaluate performance of the study's models in accurately classifying each of the 5-second audio clips. As mentioned in earlier chapters, ideally such an approach would be usable even in cases where the data exhibits an imbalance in classes (i.e. one or more classes are more prevalent than another). It should also provide a means for comparing performance of current models against historical efforts by other researchers. Schuller et al. (2012) have advocated for use of two metrics to meet these requirements: unweighted average recall, and the AUC - the area under the Receiver Operating Characteristic curve (ROC). Unweighted average recall can be used in settings where there is class imbalance, and it is the metric adopted in much of the literature treating detection of affect and social stance from paralinguistic features of speech (Schuller et al., 2012; 2013). Motivation for utilizing the AUC also derives from its extensive use automated detection of both social signals and emotion, allowing for comparison of past and current efforts (Schuller et al., 2012). In the current study,



both metrics will be calculated by applying the optimized models to the test set. Unweighted average recall and the AUC are described below.

### Unweighted Average Recall

A model's recall is defined as the proportion of true positive classifications made by a given model to the sum of its true positive (TP) *and* false negative (FN) classifications. As presented here, recall can be specified as follows:

$$Recall = \frac{TP\ X}{TP\ X + FN\ X} \quad (3.11a)$$

Calculated for high scoring audio clips, recall is the total number of high-scoring audio clips correctly identified as such (*true positives*) divided by the total number of high-scoring clips correctly identified as such *plus* the number of high scoring clips inaccurately classified as low scoring (*false negatives*). Because there are two classes of interest in the current study, recall values can be calculated for high scoring clips as well as low scoring clips. Both recall values can then be averaged, giving the unweighted average recall. Stated more formally, given two classes of observations, X and its compliment, unweighted average recall can be specified as,

$$UAR = \frac{1}{2} \left( \frac{TP\ X}{TP\ X + FN\ X} + \frac{TP\ \sim X}{TP\ \sim X + FN\ \sim X} \right) \quad (3.11b)$$

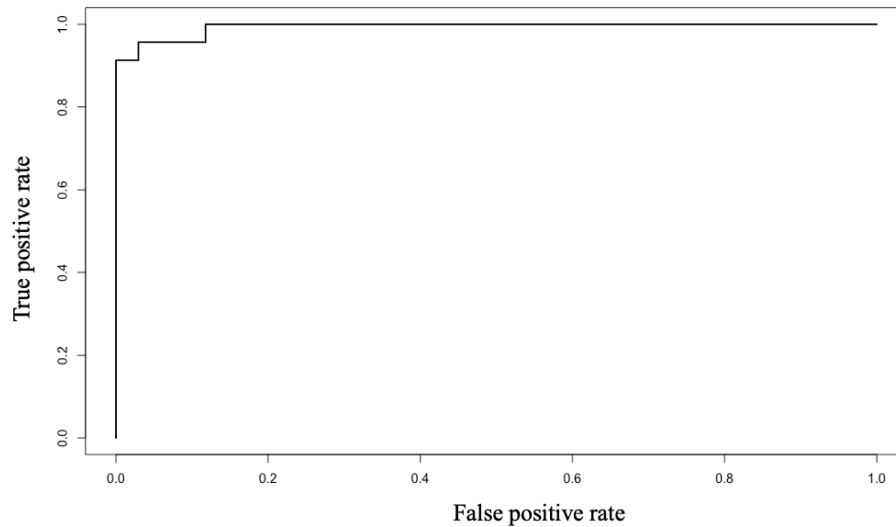
where:

- UAR is the unweighted average recall;
- TP X is the number of accurate classifications of class X made by the model;
- FN X is the number of false negative classifications of class X made by the model;
- TP ~X is the number of accurate classifications of the compliment, ~X, made by the model;
- FN ~X is the number of false negative classifications of the compliment made by the model.

### Area Under the Curve

The AUC, or area under the curve, is strictly used for binary classification problems. It is a single value indicating the area under the receiver operator curve (ROC). The ROC is a plot of the true positive rate of a model versus the false positive rate calculated for all threshold values for a model (Hajian-Tilaki, 2013; James et al., 2013). An AUC value of 0.5 indicates that a model is performing close to chance. A value of 1 indicates that the model is perfectly classifying cases, and a value of 0 indicates that it is inverting all classes. A sample receiver operator curve is presented in Figure 3.5 for reference. The ROCR package (Singh et al., 2005) is used to calculate the AUC values for each of the current study's models.

As stated previously, benchmark values for unweighted average recall and AUC employed here are inherited from work by Schuller et al. (2012) as part of the INTERSPEECH Challenge for 2012. With the intention of setting a benchmark for the field, Schuller et al. (2012) used a random forest classifier to achieve an unweighted average recall of 0.59 with an AUC of 0.647 in a binary classification task classifying speech as likable or not likable. Their results are close to but better than chance, indicating both the difficulty of the general problem of inferring social signals from acoustic features of speech and the need for continued work in this area.

**Figure 3.5***Sample Receiver Operator Curve*

*Note.* The true positive rate is also referred to as ‘sensitivity’. True positives are test or model results that correctly identify the presence of a condition or characteristic. The false positive rate is a test or model results that mistakenly identify the presence of a condition or characteristic when it is not present. The true positive rate is the proportion of true positives to the total of true positive results *and* the total of all false negative results:  $TPR = TP / TP + FN$ . The false negative rate is the proportion of false negatives to the total number of false negatives *and* true positives:  $FNR = FN / FN + TP$ .

Performances of the models will also be compared using only the acoustic feature set, the lexical feature set, and a combined set that includes both the acoustic and lexical features. Separate analyses will be carried out for the high-rated and the low-rated clips for competence-focused speech, and for the high-rated and low-rated clips for likability-focused speech. This results in a total of eighteen contrasts, as summarized in Table 3.4; model approaches (3) x feature sets (3) x constructs (competence-focused and likability-focused speech).

**Table 3.4***Summary of Comparisons Between Model Approaches, Feature Sets and Constructs*

	Feature Set 1 Acoustic Only	Feature Set 2 Lexical Only	Feature Set 3 Acoustic + Lexical
L1-Logistic Regression	CFS, LFS	CFS, LFS	CFS, LFS
Support Vector Classifier	CFS, LFS	CFS, LFS	CFS, LFS
Support Vector Machine	CFS, LFS	CFS, LFS	CFS, LFS

*Note.* CFS = classification of high and low categories of competence-focused speech; LFS = classification of high and low categories of likability-focused speech.

Benchmarks exist for detecting likability and friendliness using acoustic features of speech. Two such benchmarks are employed here to provide points of comparison, both of which were set in the 2012 INTERSPEECH challenge (Schuller et al., 2012) for detection of likability. Schuller et al. (2012) accomplished a mean unweighted average recall of 0.590 and AUC of 0.647 on the INTERSPEECH test set using a random forest model, and an unweighted average recall of 0.559 and an AUC of 0.611 using a support vector machine. The sample of recordings used in that study was a combination of scripted and unscripted speech—primarily discrete statements made in response to automated prompts from phone-based call-in system, with a mixture of call made in indoor and outdoor environments (Schuller et al., 2013).

### Results

#### FRSM Results

The mean-square residual summary statistics, infit and outfit, are presented here as indications of the fit of faceted rating scale model. Infit and outfit have an expected value of 1 and can take on values from 0 to  $+\infty$ . Mean-square values greater than 1 indicate a degree of underfit to the model, meaning the data are less predictable than the model expects. Conversely, mean-square values less than 1 indicate some degree of overfit to the model and indicate the data are more predictable than the model expects, and the scores from the rating scales tend to exhibit local dependencies. Rules of thumb for models utilizing judgements such as those made for the current study suggest that infit and outfit values less than 0.5 and greater than 2.0 degrade measurement (Gustafsson, 1980; Wright & Linacre, 1994).

#### *FRSM Performance for Competence-Focused Ratings*

Model fit statistics for rater scores of competence-focused speech are presented in Table 3.5. These results indicate that the infit and outfit values for the four rating scales for competence-focused speech fall within the desired 0.5 to 2.0 range for measurement. The ratings for intelligence and energy, however, are at the upper end of the desired range and suggest the presence of variability not captured by the rating scale model as pursued here. On the other hand, model fit statistics for overall competence indicate that ratings are somewhat more predictable than expected and suggest the presence of local dependence that may be due to the nested structure of the data.

The difficulty estimates associated with each of the scales (intelligence, motivation, energy, and overall competence) range from -0.202 to +0.611. That range in difficulty estimates suggests at least some portion of the variability in measured values for competence-focused speech originates with differences in the rating scales. The difficulty estimates for the competence rating scales suggest that it was generally easiest for raters to award higher ratings to clips for their overall competence (-0.202 logits) for example, and most difficult to award higher ratings to clips for the perceived energy (+0.611 logits) of the speakers.

**Table 3.5***Competence-Focused Speech Difficulty Estimates and Fit Statistics*

Rating scale	Difficulty (logits)	Infit MnSq	Outfit MnSq
Intelligence	-0.094	1.820	1.889
Motivation	+0.297	1.035	1.057
Energy	+0.611	1.753	1.773
Overall Competence	-0.202	0.851	0.865

*Legend.* Infit MnSq = Infit mean-square summary fit statistics; Outfit MnSq = Outfit mean-square summary fit statistics.

As noted previously, each of the rating scales uses ordinal level scores that range from 0 to 3 with higher value scores indicating increasing amounts of the property, competence-focused speech. The four possible score categories result in three possible thresholds: 0 v 1, 1 v 2, and 2 v 3. Difficulty estimates can also be identified for each of these thresholds and are provided in Appendix C2 along with additional information on the estimated facets.

Wright maps of the measured values for competence-focused and likability-focused speech were generated using the R package Wright Map (Torres & Freund, 2022). As given in Figure 10a, the Wright map relates the level of competence-focused speech exhibited by the audio-clips, the difficulty estimates for each rating scale, and raters' estimated bias, or severity, which are the three facets estimated in the faceted rating scale model, as described in Equation 1. All measured values are given on the same logit scale. In Figure 3.6a, for example, the severity of Rater 1 is greatest while the severity of Rater 5 is the smallest, indicating that Rater 5 tends to respond to audio clips with higher scores than the other raters. Conversely, Rater 1 tends to

**Figure 3.6**  
*Wright Maps for Competence and Likability Ratings*

Figure 3.6a

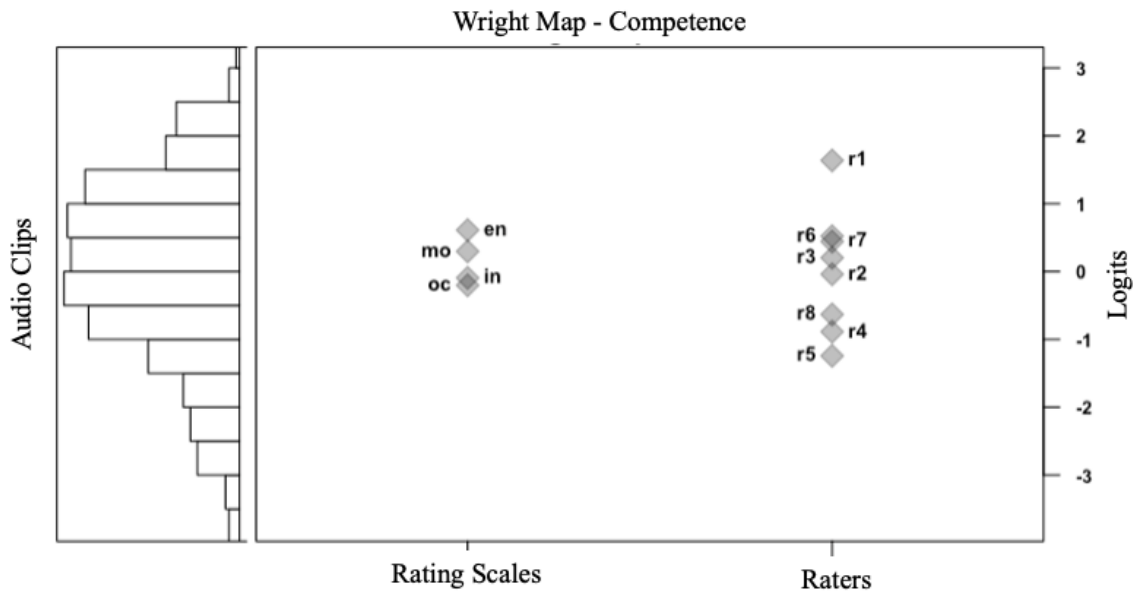
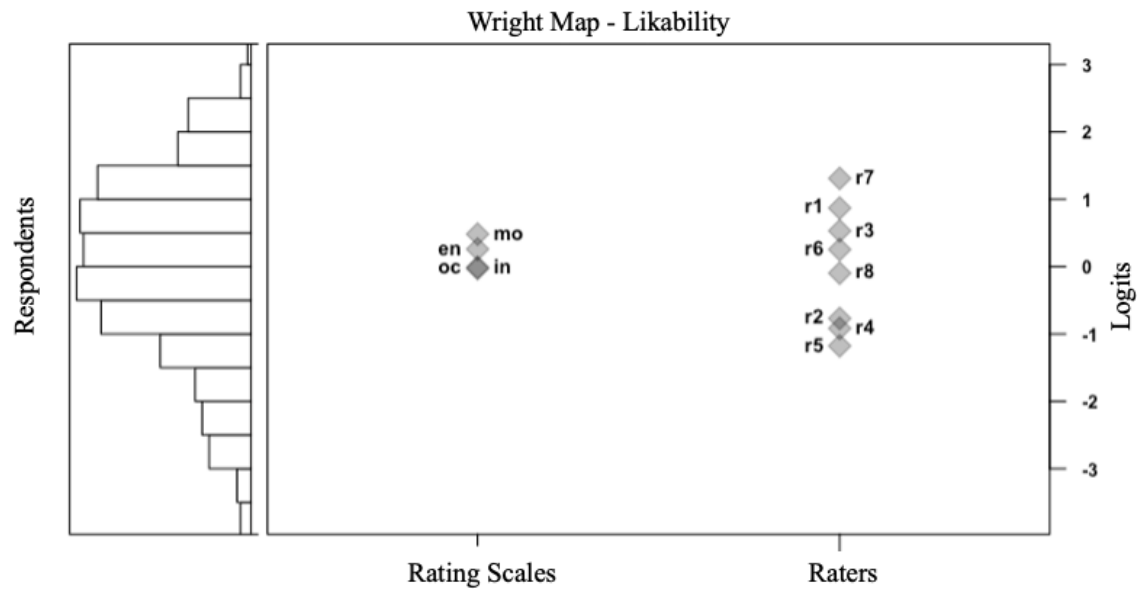


Figure 3.6b



*Legend.* en = energy, in = intelligence, mo = motivation, oc = overall competence; ca = care, fr = friendliness, wa = warmth and ol = overall likability; r1 : r8 = raters 1 through 8.

*Note.* All three facets – measured values for speech type (audio clips), difficulty of rating scales, and severity of raters are positioned vertically on the logit scale. Audio clips are ordered on the left-hand side of 5a and 5b so that clips exhibiting greater levels of the targeted speech type are higher up on the scale and those exhibiting less of the targeted speech type are located lower on the scale. Rating scales are organized vertically as well with regard to their estimated difficulty values. Raters are organized vertically with more severe raters located higher on the scale and less severe raters located lower.

respond to the same audio clips with lower scores than the other raters. Rater 5 may be thought of as a more sensitive transducer than Rater1, meaning Rater 5 tends to generate larger indication values in the presence of less of the measured property than does Rater 1.

Examination of the Wright map provides a visual demonstration of the fact that the variability in measured values of competence-focused speech for each of the audio clips is jointly influenced by the properties of the audio clips *and* the severity of the raters. The primary motivation for use of the faceted rating scale model is to account for differences in rater behavior in order to estimate measured values that are independent of those differences.

### ***Measured Values of Competence-Focused Speech Sound***

An additional motivation for using the faceted rating scale model lies in an interest in working with measured values that lie on a continuous, monotonically increasing scale as opposed to using the ordinal indication values, or rating scale scores, awarded by raters. Use of measured values placed on an interval scale makes it possible to estimate the *extent* to which the study's audio clips exhibit the targeted social signal. As a reminder, in the current chapter, the object of measurement is the speech sound recorded in each of the 5 second audio clips. The measurand, identified with the parameter  $\theta$  ('theta') in the faceted rating scale model, is the extent to which the speech sound presented in a given 5-second audio clip exhibits the targeted social signal. The unit for estimates of theta is the logit.

The values of theta for competence-focused speech had a mean of 0.010 logits (SD = 1.400) and ranged from -4.231 to +3.954. The upper bound for the first quartile of theta estimates was located at -0.827 logits. The lower bound for the fourth quartile of theta estimates was located at +0.956 logits. The lower and upper quartile values were used to identify and label the low and high scoring clips for subsequent analyses—estimation of feature importance and supervised modeling effort for competence-focused speech.

### ***FRSM Performance for Likability-Focused Speech Ratings***

Table 3.6 summarizes the infit and outfit statistics for the faceted rating scale model using rater scores for the likability-focused rating scales. As a reminder, raters used four rating scales to quantify their impressions of the speech sound in each audio clip. Raters judged the extent to which the speech sound emphasized the speaker's friendliness, warmth, care, and their overall likability. As in the case of competence-focused speech, the fit statistics for likability-focused speech fall within the desired range, 0.5 to 2.0. Fit statistics for overall likability indicate the presence of one or more local dependencies that have not been modelled. Likewise, the fit statistics for the warmth and care rating scales also suggest there may exist unmodelled dependencies. But in all cases, the fit statistics indicate the potential dependencies are not great enough to distort resulting measurements.

**Table 3.6**

#### ***Likability-Focused Speech Difficulty Estimates and Fit Statistics***

Rating scale	Difficulty (logits)	Infit MnSq	Outfit MnSq
Friendliness	-0.012	1.042	1.056
Warmth	+0.259	0.802	0.814
Care	+0.483	0.839	0.836
Overall			
Likability	-0.029	0.675	0.685

**Legend.** Infit MnSq = Infit mean-square summary fit statistics; Outfit MnSq = Outfit mean-square summary fit statistics.

The overall difficulty estimates for each of the rating scales for likability-focused speech range from -0.029 to +0.483 logits. The Wright map for likability speech is presented in Figure

3.6b and indicates that, given the dispersion of raters, variability in rater severity is a source of variability in the measured values of likability-focused speech. Use of the faceted rating scale model accounts for differences in raters' scoring behavior in order to estimate measured values that are independent of those differences. Difficulty estimates for each of the rating scale thresholds were also generated. The difficulty estimates for the rating scale thresholds as well as the other modeled facets are presented in Appendix C3.

#### ***Measured Values of Likability-Focused Speech Sound***

Theta estimates for likability-focused speech had a mean of -0.043 logits (SD = 1.049) and ranged from -2.583 to +2.654. The upper bound for the first quartile of theta estimates was located at -0.689 logits. The lower bound for the fourth quartile of theta estimates was located at +0.635 logits. These lower and upper quartile values were used to identify and label the audio clips exhibiting low and high levels of likability-focused speech for subsequent analyses.

#### **Variable Importance Results**

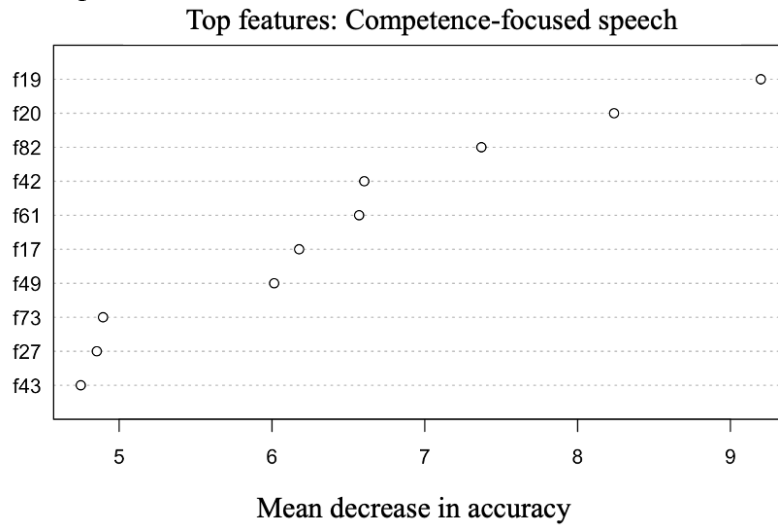
Estimation of variable importance values for each of the acoustic features allows for investigation of how speech varies between high scoring and low scoring instances of each speech type, providing some insight into how the study's models may be differentiating audio-clips in the two classification tasks. The top performing acoustic features for classifying high and low levels of competence-focused speech were identified through estimation of the mean decrease in accuracy (MDA) of the random forest model upon removal of the given feature. Examination of the top performing features reveals that the majority summarize aspects of the energy (loudness) of the speech sound or the frequencies at which the speech sound occurs.

#### ***Variable Importance for Competence-Focused Speech***

The top performing subset of the acoustic features for competence-focused speech is presented graphically in Figure 11, with their mean decrease in accuracy values. Three acoustic features related to changes in the loudness of the speech sound across the five-second clips f19, f20, and f82 have the largest impact on the mean decrease in model accuracy when classifying speech as exhibiting high or low levels of competence-focused. Examining Figure 3.7, these same three features had the largest MDA values. Features f19 and f20 summarize the rates at which speakers decrease the loudness of their speech, while f82 summarizes the number of peaks in loudness that occur per second across the audio clips, suggesting that speech sound exhibiting relatively quick decreases in volume is more likely to be perceived as exhibiting less competence-focus.

**Figure 3.7**

*Variable Importance (MDA) for Classifying Highest and Lowest Scoring Audio-Clips for Competence-Focused Speech*



*Note.* The ten acoustic features with the highest mean decrease in accuracy are listed in order from largest MDA to smaller MDAs. For descriptions of each feature see table 8 below. Note: Relative importance of features to the model accuracy is presented via estimates of the change in accuracy of the model that results from the given feature’s removal from a random forest model. MDA values are scaled by the standard deviation of the accuracy estimate. The y-axis provides the reference codes for each feature (f19, f20, etc.). These reference codes can be used with Table 3.7 to identify the feature name and description.

Table 3.7 presents the top ten features in a different format, defines each of them in order of their impact on the model accuracy, and provides the correlations between each of the features and the high and low competence-focused speech labels. The direction of the relationship between each of the acoustic features and the speech label, exhibiting *high* or *low* levels of the targeted speech type, are depicted in the final column for ease of reference.



**Table 3.7***Top Ten Performing Features for Classification of Competence-Focused Speech*

Acoustic Feature	Ref.	Category	Pt. Biserial	Relation	
<b>loudness_sma3_meanFallingSlope</b> Mean of slopes describing rate at which energy of the clip falls when it does decrease; averaged across the clip; tracks how quickly energy decreases.	f19	Energy (Loudness)	-0.587	4 <sup>th</sup> Quartile (high score) 1 <sup>st</sup> Quartile (low score)	↓ ↑
<b>loudness_sma3_stddevFallingSlope</b> Standard deviation of the rate at which loudness decreases across the audio clip; tracks how variability in rate of decreases in loudness.	f20	Energy (Loudness)	-0.569	4 <sup>th</sup> Quartile (high score) 1 <sup>st</sup> Quartile (low score)	↓ ↑
<b>loudnessPeaksPerSec.</b> Number of peaks in loudness per second; as peaks per second increases, pitch also increases.	f82	Energy (Loudness)	-0.490	4 <sup>th</sup> Quartile (high score) 1 <sup>st</sup> Quartile (low score)	↓ ↑
<b>F1bandwidth_sma3nz_amean</b> The mean difference between the upper and lower bandwidths of the first formant frequency.	f42	Frequency (Formants)	-0.387	4 <sup>th</sup> Quartile (high score) 1 <sup>st</sup> Quartile (low score)	↓ ↑
<b>hammarbergIndexV_sma3nz_amean</b> Difference in intensity of the speech sound at lower frequency bands [0-2000 Hz] and its intensity at higher bands [2000-5000 Hz]; used to judge ‘vocal effort’ of a speaker (Hammarberg et al., 1980; Schmidt, Janse et al., 2016).	f61	Spectral	+0.260	4 <sup>th</sup> Quartile (high score) 1 <sup>st</sup> Quartile (low score)	↑ ↓
<b>loudness_sma3_meanRisingSlope</b> Mean of slopes describing the rate at which energy of the clip rises when it does increase; averaged across the clip; tracks how quickly energy increases.	f17	Energy (Loudness)	-0.471	4 <sup>th</sup> Quartile (high score) 1 <sup>st</sup> Quartile (low score)	↓ ↑
<b>F2bandwidth_sma3nz_amean</b> The mean difference between the upper and lower bandwidths of the second formant frequency.	f49	Frequency (Formants)	+0.388	4 <sup>th</sup> Quartile (high score) 1 <sup>st</sup> Quartile (low score)	↑ ↓
<b>mfcc3V_sma3nz_amean</b> Mean frequency of speech sound on the Mel scale relating perceived frequency to measured frequency (voiced regions only).	f73	Frequency (MFCCs)	-0.356	4 <sup>th</sup> Quartile (high score) 1 <sup>st</sup> Quartile (low score)	↓ ↑
<b>mfcc3_sma3_amean</b> Mean frequency of speech sound on the Mel scale relating perceived frequency to measured frequency (all regions of the clip – voiced and unvoiced).	f27	Frequency (MFCCs)	-0.420	4 <sup>th</sup> Quartile (high score) 1 <sup>st</sup> Quartile (low score)	↓ ↑
<b>F1bandwidth_sma3nz_stddevNorm</b> The standard deviation of the difference between the upper and lower bandwidths of the first formant frequency.	f43	Frequency (Formants)	+0.152	4 <sup>th</sup> Quartile (high score) 1 <sup>st</sup> Quartile (low score)	↑ ↓

*Note* Here, point biserial correlations for features with the highest MDA estimates are presented alongside their relation to the outcome—high vs. low levels of *competence-focused speech*.

*Legend.* Ref = Reference code for the given acoustic feature, Pt. Biserial = Point biserial correlation.

The top performing features each fall into a set of broad categories – those relating to the variability in *energy*, or loudness of the speech sound, the *frequency* of the speech sound, and its *spectral* properties. Energy is measured in decibels and serves as an indication of perceived volume, or loudness of speech. Frequency of the speech sound is most often associated with its pitch or timbre. Spectral properties of speech combine both the energy and frequency components. Among the top ten performing features, five characterize the frequency of the clips' speech sound, four characterize energy of the speech sound, and one summarizes a spectral aspect. In what follows, the top ten performing features are grouped according to these categories and described in more detail.

***Category 1: Energy - Features summarizing variability in loudness.***

f17: Mean rate at which loudness *increases* within a given audio-clip;

f19: Mean rate at which loudness *decreases* within a given audio-clip;

f20: Standard deviation of the rate at which loudness decreases within an audio-clip;

f82: Number of peaks in loudness per second.

Features with the three highest importance values are related to the energy, or loudness, of the audio-clips' speech sound. Aspects of change in speaker energy, or loudness, are captured in three of the top performing features: f19, f20 and f82. As indicated by the correlations between labels and the feature values in Table 3.7, raters associated faster rates of change in loudness of speech (both increases and decreases) with lower levels of competence-focused speech. This is negative correlation is reflected in feature f17, which is a measure of the mean rate at which loudness increases across a given audio clip. A negative relationship is also reflected in f19, which is a measure of the mean rate at which loudness decreases across a given clip.

Variability of the rate of change, captured in f20, is gauged using the standard deviation of the slope of changes in loudness. That feature, f20, was also correlated with lower impressions of competence-focused speech. Variability in the rate at which speakers' loudness changes, is also captured by feature f82, which indicates the number of peaks in loudness per second. Increases in the number of peaks in loudness have been associated with increased pitch. Consistent with the relationship between rate of change and variability in loudness, and rater judgements of competence-focused speech, increases in the numbers of peaks in loudness per second were also negatively correlated with rater scores.

In summary then, fast increases *and* fast decreases are negatively associated with competence-focused speech, as is variability in the rate of change in loudness. In terms of everyday experiences, speakers who exhibit quick increases *or* quick decreases in loudness were less likely to be judged as emphasizing their competence. Likewise, as speakers exhibited more variability in the energy, or loudness, of their speech, they were less likely to have been perceived by the raters as emphasizing their competence.

***Category 2: Frequency - Features summarizing aspects of the formants, F1 and F2 bandwidths.***

f42: Mean difference between upper and lower bandwidths of the first formant frequency.

f49: Mean difference between upper and lower bandwidths of the second formant frequency.

f43: Standard deviation of the difference between the upper and lower bandwidths of the first formant frequency.

Descriptions of the frequency in general and the formants of speech are provided in the background section of this chapter. While the fundamental frequency, F0, is generally associated with the pitch of speakers' speech sound, and is something that listeners can easily perceive, the first and second formants are less easily associated with everyday perceived experience.

Nevertheless, in the current study, aspects of the two formant frequencies were effective in distinguishing audio-clips that judges rated as highly competence-focused from those that were rated as exhibiting very little competence-focused speech.

Feature f42, the mean bandwidth of the first formant, is negatively correlated with raters' judgements of competence-focused speech. This suggests that speech in which amplitude of the second formant has a consistently narrow range was likely to be perceived as more competence-focused. On the other hand, the mean bandwidth of the second formant, feature f49, is positively associated with judgements of competence-focused speech, suggesting that increases in range of the amplitude of the second formant was positively associated with competence-focused speech.

***Category 3: Frequency - Mel frequency cepstral coefficients (MFCC).***

f73: Mean frequency of speech sound on the Mel scale relating perceived frequency to measured frequency (voiced regions only).

f27: Mean frequency of speech sound on the Mel scale relating perceived frequency to measured frequency (all regions of the clip – voiced and unvoiced).

MFCC values indicate measured frequencies of speech sound on a scale that better reflects how speech is perceived by humans (Kent and Read, 1992). Feature f73 and f27 both indicate the mean perceived frequency of the speech sound across a given audio-clip. Feature f73 averages the MFCC values for voiced regions of the clip. Feature f27 provides the same values for both voiced and unvoiced regions of clips. In both cases, speech sounds presenting lower mean frequencies are positively correlated with higher indication scores for competence-focused speech.

***Category 4: Spectral - The Hammarberg index***

f61: The difference in energy of the speech sound at lower frequency bands [0-2000 Hz] compared to its energy at higher bands [2000-5000 Hz].

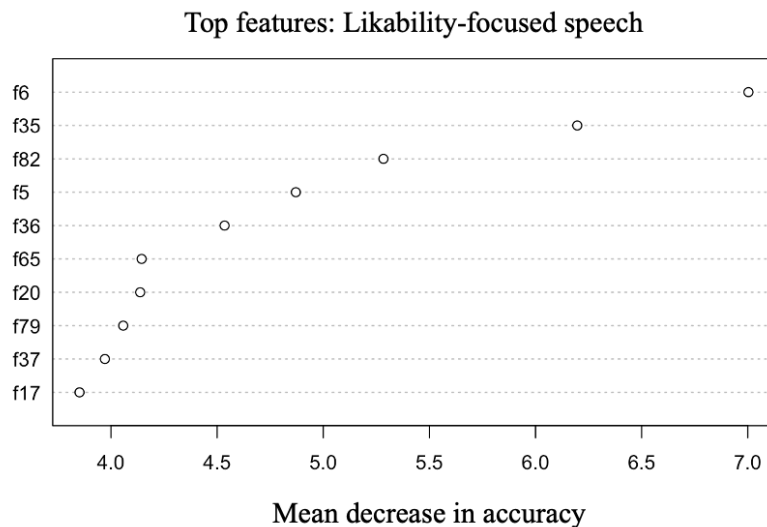
The Hammarberg Index is measured in decibels, dB, and calculated as the difference between the maxima of energy in the range of 0-2000Hz and that in the 2000-5000Hz range (Eyben, 2015). Thus, it summarizes properties of both energy and frequency of speech sound. Increased differences between energy levels of speech within the two ranges of frequencies were found to be positively correlated with raters' impressions of competence-focused speech.

***Variable Importance Results: Likability-Focused Speech***

The ten acoustic features with the highest MDA values for likability-focused speech are presented in graphically in Figure 12. Five of the highest performing acoustic features relate to energy of the speech sound, with the remaining five providing information about either frequency or spectral properties of the audio-clips. Interestingly, a subset of the top performing energy-related features is duplicated across competence-focused speech and likability-focused speech. Both relate to the rate of changes in loudness. Among the top performing acoustic features providing information about frequencies of the speech sound, three of them summarize aspects of the fundamental frequency, F0, or what can be identified with perceptions of pitch.

**Figure 3.8**

*Variable Importance (MDA) for Classifying Highest and Lowest Scoring Audio-Clips for Likability-Focused Speech*



*Note.* Relative importance of features to the model accuracy is presented via estimates of the change in accuracy of the model that results from the given feature’s removal from a random forest model. MDA values are scaled by the standard deviation of the accuracy estimate. Descriptions of each feature are given in Table 3.8.

Table 3.8 presents the top ten features for classifying likability-focused speech in a different format, defines each of the features, and provides the correlations between the features and the ‘high’ and ‘low’ competence-focused speech labels. As in the case of Table 3.7 for competence-focused speech, the direction of the relationship between each of the acoustic features and the speech label are depicted in the final column for ease of reference. For classifying likability-focused speech, the top performing features fall into a set of broad categories—those relating to the variability in *energy*, or loudness, of the speech sound, the *frequency* of the speech sound, and its *spectral* properties. In what follows, the top performing features are grouped according to these categories and described in more detail.

**Table 3.8**  
*Point Biserial Correlations for Features With the Highest MDA Estimates and Their Relation to the Outcome—Top and Bottom Quartile of Scores for Likability-Focused Speech*

Acoustic Feature	Ref.	Category	Pt. Biserial	Relation	
<b>F0semitoneFrom27.5Hz_sma3nz_pctlrange 0.2</b> Range of F0 between the 20 <sup>th</sup> and 80 <sup>th</sup> percentiles starting at 27.5Hz; higher values indicate larger changes in F0 and dynamicism in pitch.	f6	Frequency (F0)	-0.019	4 <sup>th</sup> Quartile (high score)	↓
<b>HNRdBACF_sma3nz_amean</b> Mean proportion of energy in harmonic components to energy in noise like components	f35	Energy (HNR)	-0.284	1 <sup>st</sup> Quartile (low score)	↑
<b>loudnessPeaksPerSec</b> Number of peaks in loudness per second; as peaks per second increases, pitch also increases.	f82	Energy (Loudness)	-0.330	4 <sup>th</sup> Quartile (high score)	↓
<b>F0semitoneFrom27.5Hz_sma3nz_percentile 80.0</b> Range of F0 up to the 80 <sup>th</sup> percentile starting at 27.5Hz; higher values indicate larger changes in F0 and indicates dynamicism in pitch.	f5	Frequency (F0)	-0.117	1 <sup>st</sup> Quartile (low score)	↑
<b>HNRdBACF_sma3nz_stddevNorm</b> Standard deviation of the mean proportion of energy in harmonic components to energy in noise like components.	f36	Energy (HNR)	+0.019	4 <sup>th</sup> Quartile (high score)	↑
<b>slopeV500.1500_sma3nz_amean</b> Mean spectral slope within 500-1500Hz of voiced segments of the clip; increased slope associated with perceptions of loudness and effort (Duvvuru and Erickson, 2013).	f65	Spectral	-0.157	1 <sup>st</sup> Quartile (low score)	↑
<b>loudness_sma3_stddevFallingSlope</b> Standard deviation of the rate at which loudness decreases across the audio clip; tracks how variability in rate of decreases in loudness.	f20	Energy (Loudness)	-0.284	4 <sup>th</sup> Quartile (high score)	↓
<b>slopeUV0.500_sma3nz_amean</b> Mean spectral slope within 0-500Hz of voiced segments of the clip; increased slope associated with perceptions of loudness and effort (Duvvuru and Erickson, 2013).	f79	Spectral	+0.281	1 <sup>st</sup> Quartile (low score)	↓
<b>logRelF0.H1.H2_sma3nz_amean</b> Ratio of energy of the first F0 harmonic (H1) to the energy of the second F0 harmonic (H2).	f37	Frequency (F0)	-0.120	4 <sup>th</sup> Quartile (high score)	↓
<b>loudness_sma3_meanRisingSlope</b> Mean rate at which loudness increases across the audio clip; tracks how quickly loudness increases.	f17	Energy (Loudness)	-0.286	1 <sup>st</sup> Quartile (low score)	↑

*Legend.* Ref = Reference code for the given acoustic feature, Pt. Biserial = Point biserial correlation.

*Note.* Here, point biserial correlations for features with the highest MDA estimates are presented alongside their relation to the outcome – high vs. low levels of *competence-focused speech*.

**Category 1: Energy - Features summarizing variability in loudness.**

f17: Mean rate at which loudness *increases* within a given audio-clip;

f20: Standard deviation of the rate at which loudness *decreases* within an audio-clip;

f82: Number of peaks in loudness per second.

Feature f17 summarizes the rate at which speakers’ change the loudness of their speech sound across a given audio clip. Feature f20 summarizes the variability in the rate at which speakers change the loudness of their speech across a given audio clip. Both features are negatively correlated with scores for likability-focused speech. Feature 82 is a count of the number of peaks in the energy, or loudness, of speech per second, and is also negative associated with likability-speech.

***Category 2: Energy - Harmonic to noise ratio (HNR).***

f35: Mean proportion of energy in harmonic components to energy in noise components of the speech sound of an audio clip;

f36: Standard deviation of the mean proportion of energy in harmonic components to energy in noise like components of the speech sound.

The harmonics to noise ratio (HNR) represents the ratio between periodic and non-periodic components of speech sound. Mechanically, non-periodic components of speech sound are generated at the vocal folds and caused by air escaping past the glottis (Fernandes et al., 2018). In combination, low levels of energy associated with harmonic sound and high levels energy associated with nonperiodic components of speech, i.e., noise, are associated with hoarseness (Yumoto et al., 1982). Where that is the case, results for feature f35 suggest that as speakers exhibit increasing levels of hoarseness in their speech sound, they are more likely to have been perceived as emphasizing likability-focused speech.

Feature f36 indicates that greater variability in the HNR values, as measured by the standard deviation of HNR values across a given clip are also positively correlated with higher rater scores for likability-focused speech. This feature however is difficult to relate to everyday experience, however.

***Category 3: Frequency - Fundamental Frequency***

f5: Range of F0 up to the 80<sup>th</sup> percentile starting at 27.5Hz;

f6: Range of F0 between the 20<sup>th</sup> and 80<sup>th</sup> percentiles starting at 27.5Hz;

f37: Ratio of energy of the first F0 harmonic (H1) to the energy of the second F0 harmonic (H2).

Fundamental frequency, or F0, is identified with the lowest frequency of the speech sound. As noted previously, perceptually F0 is often identified with the pitch of a voice. Features f5 and f6 both summarize information about the range of F0 in the speech sound. In both cases, increased range of F0 is negatively correlated with raters' scores of likability-focused speech.

Feature f37 provides more detailed information about the fundamental frequency. It is the ratio between the measure of energy of the first and that of the second harmonic of F0. In the current study it is negatively correlated with scores for likability-focused speech, but is also difficult to relate to the everyday experiences of human listeners.

***Category 4: Spectral features***

f65: Mean spectral slope within 500-1500Hz of voiced segments of the clip;

f79: Mean spectral slope within 0-500Hz of voiced segments of the clip.

Spectral features are properties of speech sound that combine measures of energy and frequency, providing information about how energy levels vary in relation to frequency measures. Feature f65 is a measure of the spectral slope of the speech sound—the relation between the amplitude and energy—within the 500 to 1500Hz range for voiced segments of the given audio-clip. Feature f79 presents the same measure for the 0 to 500Hz range for unvoiced segments of the audio-clip. In the case of feature f65, increased frequency of the speech sound in conjunction with level or increasing energy is negatively associated with ratings of likability-focused speech. Feature f79 on the other hand, has a positive correlation with ratings of likability-focused speech. Because both features summarize changes in the relationship between

amplitude and frequency of the speech sound across an audio clip, they too are difficult to relate to everyday experience.

### Classification Model Results

Test set results of the three different modeling approaches for classification are summarized in Table 3.9a and Table 3.9b. A total of eighteen classification models were developed, varying the construct, type of model, and the feature set by acoustic features only, lexical features only, and the combined acoustic plus lexical feature set. In each case, models were developed within a cross-validation framework with a fold size of 10.

**Table 3.9** *Model Performance Results*

Table 3.9a. Model Results for Test-Set Classification of Top and Bottom Quartiles of Scores for Competence-Focused Speech

	<u>Acoustic</u>				<u>Lexical</u>				<u>Acoustic + Lexical</u>			
	sens	spec	uar	auc	sens	spec	ua	auc	sens	spec	ua	auc
L1	0.839	0.625	0.746	0.820	0.677	0.625	0.655	0.669	0.742	0.625	0.691	0.766
SVC	0.742	0.667	0.709	0.728	0.548	0.667	0.600	0.676	0.839	0.630	0.746	0.788
SVM	0.615	0.880	0.719	0.862	0.583	0.645	0.618	0.687	0.774	0.667	0.727	0.852

*Legend.* sens: sensitivity; spec: specificity; uar: unweighted average recall; AUC: area under the curve. The positive class for the models was competence-focused speech. Low sensitivity values and high specificity indicate the models are doing a poor job accurately classifying competence-focused speech.

Table 3.9b. Model Results for Test-Set Classification of Top and Bottom Quartiles of Scores for Likability-Focused Speech

	<u>Acoustic</u>				<u>Lexical</u>				<u>Acoustic + Lexical</u>			
	sens	spec	ua	auc	sens	spec	ua	auc	sens	spec	ua	auc
L1	0.636	0.529	0.600	0.521	0.758	0.412	0.640	0.683	0.606	0.647	0.620	0.740
SVC	0.515	0.647	0.560	0.726	0.576	0.529	0.560	0.668	0.667	0.588	0.640	0.672
SVM	0.576	0.706	0.620	0.668	0.529	0.576	0.560	0.672	0.636	0.941	0.740	0.850

*Legend.* sens: sensitivity; spec: specificity; uar: unweighted average recall; AUC: area under the curve. The positive class for the models was competence-focused speech. Low sensitivity values and high specificity indicate the models are doing a poor job accurately classifying competence-focused speech.

Regarding model accuracies, the best performances were achieved by the support vector classifier and the support vector machine with radial kernel, using both sets of features. In the competence-focused speech classification task, using both the acoustic and the lexical features, the SVC had an unweighted average recall of 0.746 and an AUC of 0.788. Sensitivity was 0.839 and specificity was 0.630. In the likability-focused speech classification task, the SVM had an unweighted average recall of 0.740 and an AUC of 0.850. Sensitivity for the model was 0.636 and specificity was 0.941.

Regarding unweighted average recall, the weakest performances occurred in the context of the likability-focused data, using only a single feature set: acoustic features or lexical features only. For example, when classifying the likability-focused speech using the acoustic feature set only, the SVC had an unweighted average recall of 0.560 and an AUC of 0.726. Sensitivity was 0.515 and specificity was 0.647. Performances were similar for the SVC and the SVM when using only the lexical features.

Across the eighteen models described here, unweighted average recall of the models was improved through use of both acoustic and lexical features in most cases, but not all. In some instances, the improvement was dramatic, as in the case of the SVM classifier in the likability-classification task. In that case, unweighted average recall improved by 32%, from 0.560 using the lexical features only to 0.740 when using the combined acoustic and lexical feature sets instead of the lexical features only. The L1 logistic regression classifier, however, produced results that were inconsistent across feature sets. When classifying competence-focused speech, the L1 logistic regression classifier performed best with only acoustic features. Its performance dropped with inclusion of the composite acoustic and lexical features. When used to detect likability-focused speech, the L1 logistic regression classifier performed best when it had access to only the lexical features of speech. One possible explanation for this pattern in the performance of the L1 regression is that the use of multiple sources of information – acoustic + lexical features in this case – presents non-linear boundaries between classes where the L1 logistic classifier is designed to perform best in cases where classes are linearly separable.

### **Discussion**

Three topics for discussion are presented here. First, by identifying the acoustic features that best distinguish high scoring from low scoring audio clips, one can begin to describe the behaviors speakers may engage in when emphasizing either competence-focused or likability-focused speech. Second, development of inferential detectors (IDs) necessarily involves researchers in creating measured values of human judgements and requires a means of handling variability in the way the raters transduce one or more signals and utilize any rating scales used. Use of the faceted rating scale model is suggested as a means for identifying, quantifying and ultimately controlling for such sources of variability—here, rater severity is emphasized. Third, while there exists a great deal of important work investigating multimodal signals and how they may be leveraged for research and development efforts, composite signals may be leveraged, as well, to improve the accuracy of models for classification and prediction.

### **Speaker Behaviors**

Estimation of variable importance for acoustic features of the study's audio clips was intended to support identification of acoustic features helpful in differentiating speech exhibiting high levels of competence-focus and/or likability-focus from speech that does not. In cases where there were behavioral correlates for the features, i.e. clear explanatory links between the acoustic feature(s) and the behavior(s) that generated those features, variable importance values also suggested how speakers' behavior changes when emphasizing one or both types of speech. Such interpretations were made difficult however, where measures of acoustic properties were either indirect, involving two or more properties of the speech sound, or involved use of transformations (Fourier transforms, for example). Features involving direct measures of loudness, energy, and in some instances, pitch (F0), were an exception.

Among the top performing features, energy, pitch, and frequency all seem to play a role in identifying social signals associated with speech sounds. Regarding energy, or its perceptual correlate loudness, variation may be acceptable but rapid changes in energy seem to garner lower



impressions of both competence and likability. For example, rapid changes in loudness—either increases *or* decreases in loudness—seem to be negatively correlated with human impressions of competence-focused speech. That finding was repeated in the case of likability-focused speech, as well. There, too, quick changes in volume were negatively correlated with high ratings for likability-focused speech.

Where speakers' fundamental frequency (abbreviated as F0 here and generally associated with pitch) played an important role in classification of likability-focused speech, it was not one of the top ten features for competence-focused speech. In the case of competence-focused speech, properties of the first and second formants played important roles.

In the case of likability-focused speech, results also suggested that larger ranges in the fundamental frequency were negatively correlated with high scores from raters. Combined, these patterns in the feature-to-label correlations suggest that raters tended to award higher likability scores for speech exhibiting steady loudness and exhibiting less variability in the lowest frequency of the speech sound, the fundamental frequency. A helpful next-step may be selection of acoustic features that admit clear feature-to-behavior descriptions in order to better understand how speaker behaviors differ between high and low scoring instances of both speech types.

### **Use of the Faceted Rating Scale Model for Inferential Detectors**

A defining feature of inferential detectors is the reproduction of human judgements or perceptions even in the absence of human observers. In such a context, use of the faceted rating scale model provides a means to create measured values that meet rigorous criteria of invariant measurement; invariant in the sense that the measured values remain the same regardless of the conditions of measurement, such as the specific raters that were employed or the sample of respondents (Linacre, 2010; Engelhard, 2013; Mari & Wilson, 2014). The current study provides a successful existence proof for use of the faceted rating scale model in the context of detectors for social signals. Similar work exists in other fields that provides additional proofs for the benefits of the approach. Kennedy et al. (2020), for example, employ a version of the many facets Rasch model as a part of their process for developing an inferential detector capable of mimicking human judgements of hate speech. Use of the faceted rating scale model provides means for identifying and quantifying rater bias while also creating measured values at the interval level of measurement.

### **Leveraging Multiple Sources of Information**

At the outset of this study, it was suggested that inclusion of multiple non-redundant sources of information in the development of inferential detectors may improve the accuracy of their performance. This suggestion is supported by recent research investigating use of multimodal sources of information for detection of signals. D'Mello and Kory (2015), for example, provide an overview of this work in the context of multimodal detection of affect. Vinciarelli et al. (2008) and Vinciarelli and Esposito (2018) provide overviews of several investigations into detection of multimodal communication of social signals in general.

However, if multimodal is defined as communication occurring through more than one sensory channel (Partan & Muller, 2005), then the use of multiple modalities may be a sufficient but not a necessary requirement to produce multiple, simultaneous channels of information. Evidence in support of this view has long been present in the literature. Marler (1967) and Wickler (1978) for example, distinguish between *composite* and *multimodal* signals. On their account, multimodal signals involve more than one channel of information conveyed via two or more sensory organs. But a communicatory performance can convey information along multiple simultaneous channels using only a single modality. The phenomenon of double articulation

described here points to this possibility: even a single modality may provide multiple sources of non-redundant information. In the case of speech, for instance, it appears that there are at least two sources of information operating simultaneously, which can be identified as its segmental and suprasegmental aspects, i.e., its lexical and acoustic features.

Results from the current study support this view. But the results also suggest that there are nuances that will be important to investigate. Using the speech data and human ratings presented in the current study, model accuracies generally improved when the models included acoustic *and* lexical features of speech. In some instances, improvement of model accuracy in terms of their recall was notable, as with the use of the SVM to classify likability-focused speech where the unweighted average recall was improved by 32%. In a minority of the cases, however, model performance declined with inclusion of multiple sources of information, as observed with the L1-logistic classifier, possibly due to the introduction of a more complex feature space and less linearly separable classes.

Interestingly, the inclusion of lexical features as inputs to the study's classification models results in performances that favorably compare to Schuller et al. (2012) benchmarks as described in the methods section of the current chapter. In that work, Schuller et al. accomplished a mean unweighted average recall of 0.590 and AUC of 0.647 on the INTERSPEECH test set using a random forest model, and an unweighted average recall of 0.559 and an AUC of 0.611 using a support vector machine. In the current work, the support vector machine utilizing acoustic and lexical features had an unweighted accuracy of 0.740 and an AUC of 0.850.

### Future Work

The accuracy of the models was generally improved in most but not all cases with use of the composite feature set (acoustic features plus lexical features). This underscores the fact that there are many degrees of freedom in development of inferential detectors for social signals. A full understanding of the benefits of multiple sources of information for inferential detection will necessarily require investigation of the effects of several variables beyond simply the *number* of sources of information used. Researchers will have choices **among** the various sources of information used in the modelling process, how those features are summarized or represented in the model(s), and which model(s) are employed. In the current study for instance, lexical features of speech were represented within the modeling process through estimations of the mutual information associated with each word. But there are many approaches to representing propositional content and these should be explored (see Bengio et al., 2013 for an example). Likewise, several researchers have begun identifying and investigating a variety of approaches for combining, or fusing, information from multiple sources into the modeling process (Attrey et al., 2010; Poria et al., 2017, e.g.) and the benefits of each under which conditions will need to be investigated, as well. Likewise, as evidenced here, it will be valuable to understand how the performance of different modeling approaches changes with inclusion of composite and multimodal sources of information.

### Conclusion

Linear measures that account for bias in human judgements can be created and leveraged for classification of social signals according to human impressions. The faceted rating scale model can play an important role in identifying and accounting for differences in the ways individual raters interpret one or more rating scales. Importantly, resulting human impressions of speech can be inferred with moderate levels of accuracy. The current work provides an existence proof for detection of social signals even when using extemporaneous speech that occurs in

naturalistic environments. Inclusion of multiple sources of information from composite signals, in this case lexical as well as acoustic features of speech, stands to improve the accuracy with which inferential detectors mimic the impressions of human observers.

**Appendix A.1: R Scripts for Chapter 1**

```
#FULL PIPELINE
#ti: Load Data
#au: smc
#cr: 01-01-2022
#mod: 04-27-2022
```

```
#Load joined expert file
expert.all <-
read.csv("/Users/s.corrigan/Desktop/Dissertation/6.Dissertation_Data/DissertationData_Mar30_2022/g2.expert.all.tree.csv")
expert.all$spch_type <- as.factor(expert.all$spch_type)
levels(expert.all$spch_type) <- c("C", "L")
```

---

```
#ti: Check_missing and outliers
#au: smc
#cr: 02-01-2022
#mod: 04-27-2022
```

```
#Check for missing data
sapply(expert.all, function(x) sum(is.na(x)))
sapply(nonexpert.all, function(x) sum(is.na(x)))
```

```
#Standard and center
library(tibble) #using tibble to reinsert spch_type after scaling
str(expert.all) #Confirms features are in positions 3:90
expert.all.z <- as.data.frame(scale(expert.all[,3:90]))
expert.all.z <- add_column(expert.all.z, expert.all$spch_type, .before = "f1")
colnames(expert.all.z)[1] <- 'spch_type'
expert.all.z <- add_column(expert.all.z, expert.all$userId, .before = "spch_type")
colnames(expert.all.z)[1] <- 'userId'
str(expert.all.z)
```

```
str(nonexpert.all) #Confirms features are in positions 3:90
nonexpert.all.z <- as.data.frame(scale(nonexpert.all[,3:90]))
nonexpert.all.z <- add_column(nonexpert.all.z, nonexpert.all$spch_type, .before = "f1")
colnames(nonexpert.all.z)[1] <- 'spch_type'
nonexpert.all.z <- add_column(nonexpert.all.z, nonexpert.all$userId, .before = "spch_type")
colnames(nonexpert.all.z)[1] <- 'userId'
str(nonexpert.all.z)
```

```
#Initial check for outliers
```

```

library(tidyverse)

expert.all.z %>%
  select(f2,f3,f4,f5,f6,f7,f8,f9,f10) %>%
  map_df(.f = ~ broom::tidy(summary(.x)), .id = "variable")

boxplot.stats(expert.all.z$f2)$out #identify outlier values
out <- boxplot.stats(expert.all.z$f2)$out #identify associated row numbers of outliers
out_ind <- which(expert.all.z$f2 %in% c(out))
out_ind
#install.packages("mvoutlier")
library(mvoutlier)
expert.sub = expert.all.z[, 1:3] #first three variables
result = mvOutlier(expert.sub, qqplot = TRUE, method = "quan", label = TRUE)
result

```

```

#Check balance across the whole set
expert_total <- nrow(expert.all.z)
expertC_count <- sum(with(expert.all.z, spch_type == "C"))
expertL_count <- sum(with(expert.all.z, spch_type == "L"))
expertC_count / expert_total #0.4763729 are C
expertL_count / expert_total #0.5236271 are L

```

```

nonexpert_total <- nrow(nonexpert.all.z)
nonexpertC_count <- sum(with(nonexpert.all.z, spch_type == "C"))
nonexpertL_count <- sum(with(nonexpert.all.z, spch_type == "L"))
nonexpertC_count / nonexpert_total #0.4884772 are C
nonexpertL_count / nonexpert_total #0.5115228 are L

```

---

```

#ti: Test-Train_Split
#au: smc
#cr: 3-22-2022
#mod: 3-25-2022

```

```

#install.packages("splitTools")
library(splitTools)

```

```

#Dataframes created: 1) expert.train, 2) expert.test, 3) expert.train.labels
#Partition expert data
expert.ids <- splitTools::partition(
  expert.all.z$userId,
  p = c(train = 0.7, test = 0.3),
  type = "grouped"
)

```

```

expert.train <- expert.all.z[expert.ids$train, ]
ncol(expert.train)
drop <- c("userId")
train = train[,!(names(train) %in% drop)]
expert.train = expert.train[,!(names(expert.train) %in% drop)]
expert.train.labels <- expert.train$spch_type
nrow(expert.train) #n = 1,623 audio clips

```

```

expert.test <- expert.all.z[expert.ids$test, ]
expert.test = expert.test[,!(names(expert.test) %in% drop)]
expert.test.labels <- expert.test$spch_type
nrow(expert.test) #n = 726 audio clips

```

```

#ti: D. ModelDev_IMPORTANCEImportance
#au: smc
#cr: 4-01-2022
#mod: 5-02-2022
#https://rpubs.com/phamdinhkhanh/389752

```

```

library(caret)
expert.train.rf <- expert.train

```

```

#Create control function for training with 10 folds and keep 3 folds for training. search method is
grid.

```

```

control <- trainControl(method='repeatedcv',
  number=10,
  repeats=3,
  search='grid')

```

```

#create tuneGrid with 15 values from 1:15 for mtry to tuning model. This train function will
change number of entry variable at each split according to tuneGrid.

```

```

tuneGrid <- expand.grid(.mtry = (1:15))

```

```

rf_gridsearch <- train(spch_type ~ .,
  data = expert.train.rf,
  method = 'rf',
  metric = 'Accuracy',
  tuneGrid = tuneGrid)
print(rf_gridsearch)
rfImp <- varImp(rf_gridsearch, scale = FALSE)
str(rfImp$importance)
rfImp
library(dplyr)

```

```

col_index <- varImp(rf_gridsearch)$importance %>%
mutate(names=row.names(.)) %>%

```

```
arrange(-Overall)
imp_names <- col_index$names[1:15]
imp_names
plot(subset(rfImp, imp_names))
```

---

```
#ti: LASSOmodelDevelopment
#au: smc
#cr: 04-01-2022
#mod: 04-27-2022
```

```
#Inherit Data: expert.train,expert.test and nonexpert.train, nonexpert.test
library(caret)
library(gbm)
```

```
#Specify crossvalidation framdework for tuning
fitControl <- trainControl(## 10-fold CV
  method = "repeatedcv",
  number = 10,
  repeats = 10,
  search = 'random')
```

```
#Lasso Logistic Regression
library(caret)
library(glmnet)
str(expert.train)
#Prepare outcome variables for glmnet
expert.train.L1 <- expert.train
expert.test.L1 <- expert.test
expert.train.L1$spch_type <- as.numeric(expert.train.L1$spch_type)
expert.train.L1$spch_type <- (expert.train.L1$spch_type-1)
expert.test.L1$spch_type <- as.numeric(expert.test.L1$spch_type)
expert.test.L1$spch_type <- (expert.test.L1$spch_type-1)
str(expert.train.L1)
```

```
#Regularization with the Lasso
library(glmnet)
expert.train.matrix <- model.matrix(spch_type ~., expert.train.L1)[-1]
lambdas <- 10^seq(8, -4, length = 250)

expert.train.lasso <- glmnet(expert.train.matrix,
  expert.train.L1$spch_type, alpha = 1, lambda = lambdas, family = "binomial")
expert.train.lasso.cv <- cv.glmnet(expert.train.matrix,
  expert.train.L1$spch_type, alpha = 1, lambda = lambdas, family = "binomial")
expert.train.lasso.cv
```

```
expert.train.lasso
lambda.lasso <- lasso.cv$lambda.min
lambda.lasso
```

```
#Aside: identify features whose coefficients drive to zero and those that remain
#Remaining coefficients deserve explanation in substantive terms wrt voice quality
predict(expert.train.lasso, type = "coefficients", s = lambda.lasso)
```

```
#Apply discovered lambda to fit an optimal model
expert.train.optlasso <- glmnet(expert.train.matrix,
  expert.train.L1$spch_type, alpha = 1, lambda = 0.0164766, family = "binomial")
expert.train.lasso.probability = predict(expert.train.optlasso, expert.train.matrix, type =
"response")
expert.train.lasso.class = expert.lasso.pred > 0.5
```

```
#Outcome count
summary(expert.train.lasso.class)
```

---

```
#ti: LASSOModelDev_Eval
#au: smc
#cr: 4-1-2022
#mod: 5-02-2022
```

```
#Confusion matrix
expert.lasso.CM = table(expert.train.L1$spch_type, expert.train.lasso.class)
expert.lasso.CM
#Evaluations for train set
err_metric(expert.lasso.CM)
roc_score.train.lasso=roc(as.numeric(expert.train.L1[,1]),
as.numeric(expert.train.lasso.probability)) #AUC score
roc_score.train.lasso
```

```
#Lasso: Testset performance
expert.test.matrix <- model.matrix(spch_type ~., expert.test.L1)[-1]
expert.lasso.test.probability <- predict(expert.train.optlasso, s = lambda.lasso, newx =
expert.test.matrix, type = "response")
expert.lasso.test.class <- as.numeric(expert.lasso.test.probability > 0.5)
mean(expert.lasso.test.class == as.numeric(expert.test.L1$spch_type)) #0.6419919
```

```
#Confusion matrix
expert.lasso.test.CM = table(expert.test.L1$spch_type, expert.lasso.test.class)
expert.lasso.test.CM

err_metric(expert.lasso.test.CM )
```



```
roc_score.test.lasso=roc(as.numeric(expert.test.L1[,1]), as.numeric(lasso.test.probability)) #AUC
score
roc_score.test.lasso #0.5672
```

---

```
#ti: ModelDev_SVM - Suport Vector Classifier
#au: smc
#cr: 04-01-2022
#mod: 05-02-2022
```

```
#Linear support vector classifier
```

```
seed(2343)
```

```
expert.train$spch_type <- as.factor(expert.train$spch_type)
```

```
expert.test$spch_type <- as.factor(expert.test$spch_type)
```

```
expert.train.tune.svm.linear <- tune(svm, spch_type ~., data = expert.train, kernel = "linear",
  ranges = list(cost = c(.01, .1, 1, 10, 100)))
```

```
expert.train.tune.svm.linear
```

```
plot(expert.train.tune.svm.linear)
```

```
#Best performing parameters
```

```
expert.train.tune.svm.linear$best.parameters
```

```
#Best performance
```

```
1 - expert.train.tune.svm.linear$best.performance
```

```
#Radial kernel use 10-fold CV (default)
```

```
str(expert.test)
```

```
set.seed(2434)
```

```
expert.train.tune.svm.radial <- tune(svm, spch_type ~., data = expert.train, kernel = "radial",
  ranges = list(cost = c(.01, .1, 1, 10, 100),
  gamma = c(.01, .05, 0.1, 0.5, 1)))
```

```
expert.train.tune.svm.radial
```

```
#Best performing parameters
```

```
expert.train.tune.svm.radial$best.parameters # best: c = 10; gamma = 0.01
```

```
#Best performance
```

```
1 - expert.train.tune.svm.radial$best.performance
```

---

```
#ti: ModelDev_Eval_SVM
```

```
#au: smc
```

```
#cr: 4-1-2022
```

```
#mod: 4-28-2022
```

```
#error metrics -- Confusion Matrix
```

```
err_metric=function(CM)
{
  TN =CM[1,1]
  TP =CM[2,2]
  FP =CM[1,2]
  FN =CM[2,1]
  precision =(TP)/(TP+FP)
  recall_score =(FP)/(FP+TN)
  fl_score=2*((precision*recall_score)/(precision+recall_score))
  accuracy_model =(TP+TN)/(TP+TN+FP+FN)
  False_positive_rate =(FP)/(FP+TN)
  False_negative_rate =(FN)/(FN+TP)
  print(paste("Precision value of the model: ",round(precision,2)))
  print(paste("Accuracy of the model: ",round(accuracy_model,2)))
  print(paste("Recall value of the model: ",round(recall_score,2)))
  print(paste("False Positive rate of the model: ",round(False_positive_rate,2)))
  print(paste("False Negative rate of the model: ",round(False_negative_rate,2)))
  print(paste("f1 score of the model: ",round(fl_score,2)))
}
```

```
#ROC curve function per James et al.,2013
```

```
library(ROCR)
rocplot = function(pred, truth, ...){
  predob = prediction(pred, truth)
  perf = performance(predob, "tpr", "fpr")
  plot(perf, ...)}
```

```
#Best performance - radial svm
```

```
1 - expert.train.tune.svm.radial$best.performance #0.392861; i.e. accuracy = 0.607139
```

```
#Best performance - radial svm
```

```
expert.train.tune.svm.radial$best.model
```

```
#TRAIN: EVALUATION
```

```
#Obtain fitted values for each observation
```

```
set.seed(2356)
```

```
svmfit.optimal = svm(spch_type ~., data = expert.train, kernel = "radial",
  gamma = .01, cost = 10, decision.values = TRUE, probability = TRUE)
```

```
#see p.365 James, Witten et al.2013
```

```
fitted.train.decisionvalues = as.numeric(attributes(predict(svmfit.opt, expert.train,
  decision.values = TRUE, probability = TRUE))$decision.values)
```

```
fitted.train.decisionvalues
```

```
#TRAIN: apply prediction function on train data
```

```
train.predictions.classes <- as.numeric(predict(svmfit.optimal, expert.train[,2:89]))
```

```
#TRAIN: Confusion Matrix using CM Function (below)
```

```
CM.train= table(expert.train.auc[,1] , train.predictions.classes)
```

```
print(CM.train)
```

```
err_metric(CM.train)
```

```
#TRAIN: AUC
```

```
roc_score.train=roc(expert.train[,1], fitted.train.decisionvalues) #AUC score
```

```
roc_score.train
```

```
#TRAIN: create roc curve for training data
```

```
library(pROC)
```

```
par(mfrow=c(1,2))
```

```
rocplot(fitted.train.decisionvalues, expert.train$spch_type, main = "Training Data", auc = TRUE)
```

```
#TEST: Apply best model to test set
```

```
#Get predicted classes for confusion matrix
```

```
test.predictions.classes <- as.numeric(predict(svmfit.optimal, expert.test))
```

```
test.predictions.classes
```

```
summary(test.predictions.classes)
```

```
#Get predicted decision values for AUC
```

```
fitted.test.decisionvalues = as.numeric(predict(svmfit.opt, expert.test))
```

```
fitted.test.decisionvalues
```

```
expert.test[,1]
```

```
mean(test.predictions.classes == as.numeric(expert.test[,1])) #accuracy on test set = 0.5757576
```

```
table(predicted = test.predictions.classes, actual = as.numeric(expert.test[,1]))
```

```
#TEST: Confusion Matrix
```

```
expert.test.auc <- expert.test
```

```
expert.test.auc$spch_type <- as.numeric(expert.test.auc$spch_type)
```

```
CM.test = table(expert.test.auc[,1] , predict.test.predictedclasses)
```

```
print(CM.test)
```

```
err_metric(CM.test)
```

```
roc_score.test=roc(expert.test[,1], predict.test.decisionvalues) #AUC score
```

```
roc_score.test
```

```
# Create roc curve for test data
```

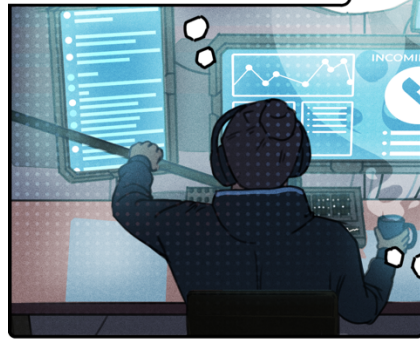
```
library(pROC)
```

```
par(mfrow=c(1,2))
```

```
rocplot(fitted.test.decisionvalues, expert.test$spch_type, main = "Test Data")
```

### Appendix A.2: Sample Recording Tasks

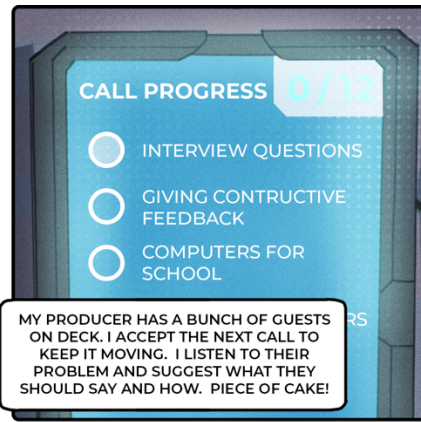
COMPUTER ON, MIC STANDING BY, COFFEE IN HAND... IT'S SHOWTIME! I LOVE MY JOB, HELPING PEOPLE SOLVE PROBLEMS AND IMPROVE THEIR COMMUNICATION SKILLS.



MY PRODUCER HAS A BUNCH OF GUESTS ON DECK. I ACCEPT THE NEXT CALL TO KEEP IT MOVING. I LISTEN TO THEIR PROBLEM AND SUGGEST WHAT THEY SHOULD SAY AND HOW. PIECE OF CAKE!

CALL PROGRESS 0/12

- INTERVIEW QUESTIONS
- GIVING CONSTRUCTIVE FEEDBACK
- COMPUTERS FOR SCHOOL

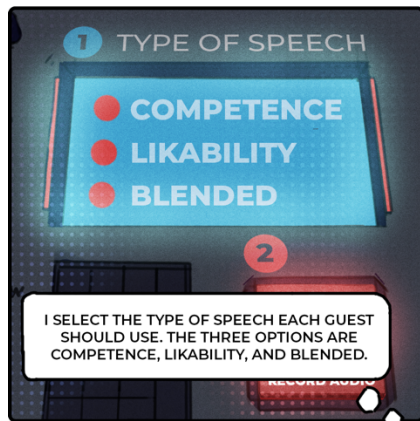


1 TYPE OF SPEECH

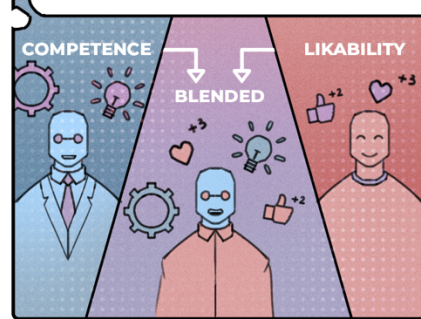
- COMPETENCE
- LIKABILITY
- BLENDED

2

I SELECT THE TYPE OF SPEECH EACH GUEST SHOULD USE. THE THREE OPTIONS ARE COMPETENCE, LIKABILITY, AND BLENDED.



COMPETENCE MEANS OTHERS PERCEIVE YOU AS INTELLIGENT, MOTIVATED, AND ENERGETIC. LIKABILITY MEANS OTHERS PERCEIVE YOU AS FRIENDLY, CARING, AND WARM. BLENDED MEANS ALL OF THE ABOVE!



MY PRODUCER WANTS ME TO PRACTICE SOUNDING AS LIKEABLE AS POSSIBLE BY RECORDING A SAMPLE INTRO:

"WELCOME TO ANOTHER EDITION OF THE ADVICE HOUR"


● BLENDED

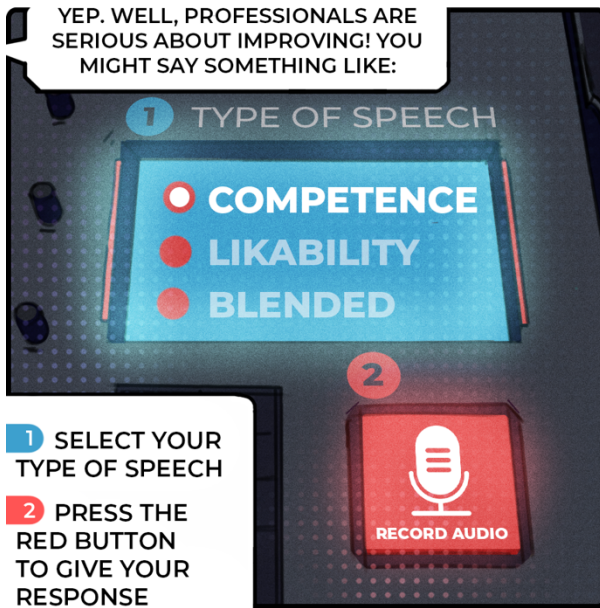
2

1 PRESS THE RED BUTTON TO GIVE YOUR RESPONSE

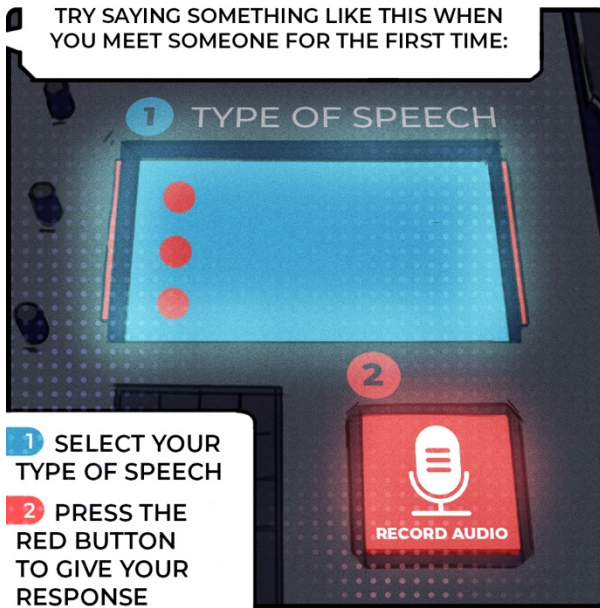
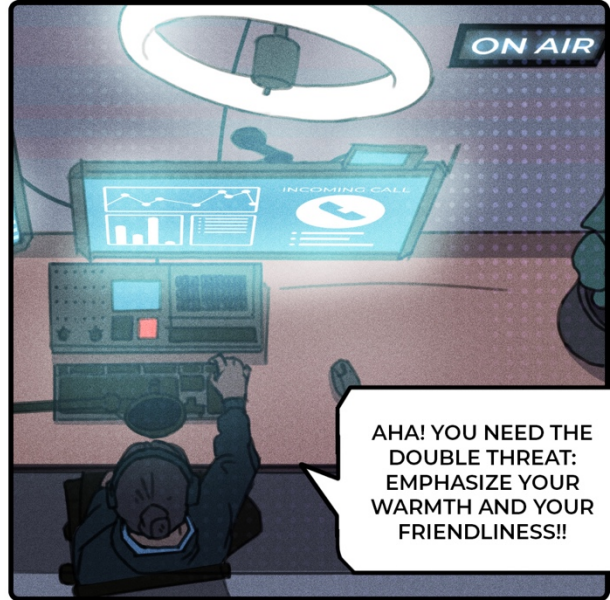
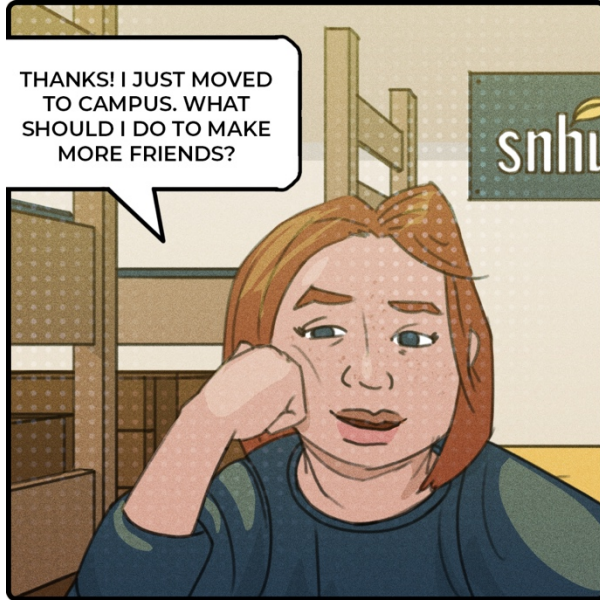


GREAT, LOOKS LIKE EVERYTHING IS GOOD TO GO! I'VE GOT A BUSY SCHEDULE TO GET THROUGH TODAY...

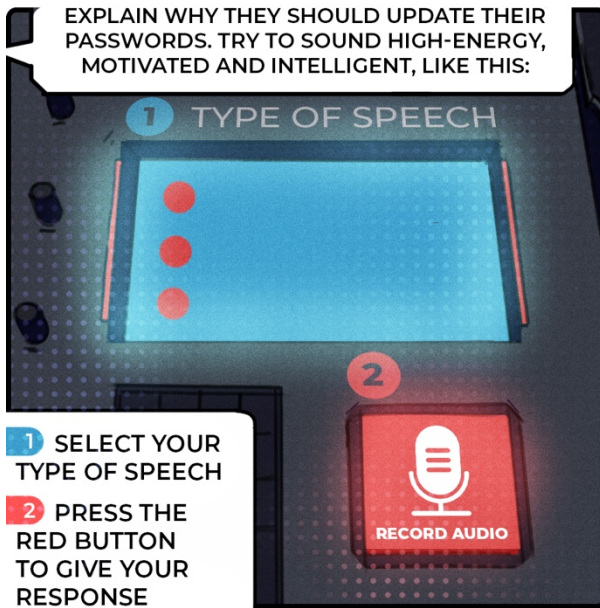












### Appendix A.3: eGeMAPS Acoustic Parameters

(Taken from Eyben et al. (2009) with some additions from Kent and Read (2002) to describe perceptual correlates where possible.)

No.	MDA	Feature Name	Description
1	2.060	F0semitoneFrom27.5Hz_sma3nz_amean	<b>Frequency Related Parameters</b> F0 semitone parameters capture information about speakers' pitch, its average value and its change over time. A semitone frequency scale is used which has its origin at 27.5Hz.
2	4.455	F0semitoneFrom27.5Hz_sma3nz_stddevNorm	
3	2.805	F0semitoneFrom27.5Hz_sma3nz_percentile20.0	
4	4.873	F0semitoneFrom27.5Hz_sma3nz_percentile50.0	
5	7.030	F0semitoneFrom27.5Hz_sma3nz_percentile80.0	
6	5.211	F0semitoneFrom27.5Hz_sma3nz_pctrange02.	
7	1.654	F0semitoneFrom27.5Hz_sma3nz_meanRisingSlope.	
8	0.198	F0semitoneFrom27.5Hz_sma3nz_stddevRisingSlope.	
9	1.377	F0semitoneFrom27.5Hz_sma3nz_meanFallingSlope.	
10	0.713	F0semitoneFrom27.5Hz_sma3nz_stddevFallingSlope.	
11	0.667	loudness_sma3_amean	<b>Energy/Amplitude Related Parameters</b> Estimates of perceived signal intensity that are expected to be perceived as the volume or loudness of speech as well as variability in loudness across a given 5-second window. Values for standard deviations and percentiles are provided.
12	0.527	loudness_sma3_stddevNorm	
13	2.569	loudness_sma3_percentile20.0	
14	2.609	loudness_sma3_percentile50.0	
15	2.024	loudness_sma3_percentile80.0	
16	4.083	loudness_sma3_pctrange02	
17	2.985	loudness_sma3_meanRisingSlope	
18	0.726	loudness_sma3_stddevRisingSlope	
19	0.630	loudness_sma3_meanFallingSlope	
20	3.580	loudness_sma3_stddevFallingSlope	
21	2.364	spectralFlux_sma3_amean	<b>Extended Parameter Set</b> Indicates the difference of the spectra across two consecutive frames.
22	3.095	spectralFlux_sma3_stddevNorm	
23	2.088	mfcc1_sma3_amean	<b>Spectral (balance/shape/dynamics) Parameters</b> Mean and Coefficients of variation for Mel-Frequency Cepstral Coefficients 1-4 across <i>all</i> regions of the recording.
24	1.721	mfcc1_sma3_stddevNorm	
25	-0.613	mfcc2_sma3_amean	
26	-0.751	mfcc2_sma3_stddevNorm	
27	2.038	mfcc3_sma3_amean	
28	3.076	mfcc3_sma3_stddevNorm	
29	1.472	mfcc4_sma3_amean	
30	0.644	mfcc4_sma3_stddevNorm	
31	1.043	jitterLocal_sma3nz_amean	
32	5.331	jitterLocal_sma3nz_stddevNorm	
33	-1.016	shimmerLocaldB_sma3nz_amean	<b>Energy/Amplitude Related Parameters</b> Capture differences in the peak amplitudes of consecutive F0 periods. Perceptually, low shimmer estimates indicate speech is produced at a stable volume or amplitude, i.e. loudness. Higher shimmer values indicate variability in the loudness of the speech.
34	1.197	shimmerLocaldB_sma3nz_stddevNorm	
35	5.190	HNRdBACF_sma3nz_amean	<b>Energy / Amplitude</b> Harmonics to noise ratio (HNR) is the relation of the energy of the harmonic components of speech to the energy of the noise-like components.
36	1.830	HNRdBACF_sma3nz_stddevNorm	
37	2.167	logRelF0H1H2_sma3nz_amean	<b>Spectral (Balance) Parameters</b> <i>Harmonic difference H1-H2</i> is the ratio of energy of the first F0 harmonic(H1) to the energy of second F0 harmonic (H2).
38	2.581	logRelF0H1H2_sma3nz_stddevNorm	
39	3.664	logRelF0H1A3_sma3nz_amean	
40	5.843	logRelF0H1A3_sma3nz_stddevNorm	



			<i>Harmonic difference H1-A3</i> is the ratio of energy of the first F0 harmonic to the energy of the highest harmonic in the third formant range (A3).
41	3.824	F1frequency_sma3nz_amean	<p><b>Frequency Related Parameters</b> F1, F2 and F3 frequency parameters describe values associated with the first, second and third formants, respectively. Formants capture information about the resonance of sound waves along the vocal tract. Lingual, nasal and labial components of the speech tract, for example, can all be manipulated to change aspect of the formant frequencies (amplitude, period) of speech to produce a range of vowel sounds, e.g. (Kent and Read, 2002).</p>
42	1.434	F1frequency_sma3nz_stddevNorm	
43	2.215	F1bandwidth_sma3nz_amean	
44	0.567	F1bandwidth_sma3nz_stddevNorm	
45	3.323	F1amplitudeLogRelF0_sma3nz_amean	
46	3.916	F1amplitudeLogRelF0_sma3nz_stddevNorm	
47	3.549	F2frequency_sma3nz_amean	
48	0.297	F2frequency_sma3nz_stddevNorm	
49	0.449	F2bandwidth_sma3nz_amean	
50	2.864	F2bandwidth_sma3nz_stddevNorm	
51	3.192	F2amplitudeLogRelF0_sma3nz_amean	
52	3.670	F2amplitudeLogRelF0_sma3nz_stddevNorm	
53	2.330	F3frequency_sma3nz_amean	
54	0.492	F3frequency_sma3nz_stddevNorm	
55	0.814	F3bandwidth_sma3nz_amean	
56	1.414	F3bandwidth_sma3nz_stddevNorm	
57	3.799	F3amplitudeLogRelF0_sma3nz_amean	
58	3.488	F3amplitudeLogRelF0_sma3nz_stddevNorm	
59	1.123	alphaRatioV_sma3nz_amean	<p><b>Spectral (Balance) Parameters</b> Alpha ratio is the ratio of summed energy from 50-1000Hz and 1-5kHz</p>
60	1.454	alphaRatioV_sma3nz_stddevNorm	
61	0.712	hammarbergIndexV_sma3nz_amean	<p><b>Spectral (Balance) Parameters</b> The ratio of the strongest energy peak in the 0-2kHz region to the strongest peak in the 2-5kHz region.</p>
62	3.237	hammarbergIndexV_sma3nz_stddevNorm	
63	4.084	slopeV0500_sma3nz_amean	<p><b>Spectral (Balance) Parameters</b> Arithmetic mean of the spectral slope across all voiced regions of the given recording for ranges 0-500 Hz and 500 to 1500Hz.</p>
64	3.956	slopeV0500_sma3nz_stddevNorm	
65	1.563	slopeV5001500_sma3nz_amean	
66	3.513	slopeV5001500_sma3nz_stddevNorm	
67	2.261	spectralFluxV_sma3nz_amean	<p><b>Extended Parameter Set</b> Spectral flux parameters indicate the of the spectra across two consecutive frames.</p>
68	-0.587	spectralFluxV_sma3nz_stddevNorm	
69	2.120	mfcc1V_sma3nz_amean	<p><b>Spectral (balance/shape/dynamics) Parameters</b> The means and coefficients of variation of Mel-Frequency Cepstral Coefficients 1-4 in voiced regions of the recording.</p>
70	6.465	mfcc1V_sma3nz_stddevNorm	
71	0.581	mfcc2V_sma3nz_amean	
72	0.133	mfcc2V_sma3nz_stddevNorm	
73	1.279	mfcc3V_sma3nz_amean	
74	1.701	mfcc3V_sma3nz_stddevNorm	
75	1.989	mfcc4V_sma3nz_amean	
76	-1.548	mfcc4V_sma3nz_stddevNorm	
77	1.838	alphaRatioUV_sma3nz_amean	<p><b>Energy/Amplitude</b> ratio of the summed energy from 50–1000 Hz and 1–5 kHz across unvoiced regions of the recording.</p>
78	2.363	hammarbergIndexUV_sma3nz_amean	
79	0.747	slopeUV0500_sma3nz_amean	<p><b>Spectral (Balance) Parameters</b> Arithmetic mean of the spectral slope across all unvoiced regions of the given recording for ranges 0-500 Hz and 500 to 1500Hz..</p>
80	2.100	slopeUV5001500_sma3nz_amean	
81	2.633	spectralFluxUV_sma3nz_amean	<p><b>Spectral (Balance) Parameters</b> Mean difference of the spectra across consecutive frames.</p>
82	2.384	loudnessPeaksPerSec	<p><b>Temporal Features</b></p>
83	1.385	VoicedSegmentsPerSec	

84	1.751	MeanVoicedSegmentLengthSec	<i>Rate of loudness peaks</i> - the number of loudness peaks occurring on average each second. <i>Mean length and SD</i> of continuously voiced regions (i.e. where $F_0 > 0$ ). <i>Mean length and SD</i> of unvoiced regions (i.e. where $F_0 = 0$ ) therefore approximating pauses in speech. <i>Number of continuous voiced regions per second</i> , indicating the syllable rate.
85	0.079	StddevVoicedSegmentLengthSec	
86	4.587	MeanUnvoicedSegmentLength	
87	1.333	StddevUnvoicedSegmentLength	
88	3.001	equivalentSoundLevel_dBp	

## Appendix B: R-Scripts for Chapter 2

### R-Scripts – Data Preparation

```

---
title: "R Notebook: Chp2.JoinFiles"
output: html_notebook
au: smc
---
A. Load files: a) base_data, b) spch_type labels, c) demographics
```{r}
wd <- getwd()
list.files(wd)
base_data <- read.csv("/cloud/project/base_data.csv")
demo <- read.csv("/cloud/project/demographics.csv")
egemaps <- ("/cloud/project/opensmile_egemaps.csv")
compare <- read.csv("/cloud/project/opensmile_compare.csv")
module <- read.csv("/cloud/project/module_lookup.csv")
```

B.1. Initial Cleaning: Change acoustic feature names to fl:f6374
```{r}
recordid <- as.data.frame(compare$f1)
str(recordid)
colnames(recordid) <- c("recordid")
str(recordid)
drop1 <- c("f1")
compare = compare[!(names(compare) %in% drop1)]
head(compare)
colnames(compare) <- paste0("f",1:ncol(compare))
head(compare)
#library(tibble)
compare <- add_column(compare, recordid, .before = 1) # Apply add_column function by index
head(compare)
ncol(compare)
```

B2. Initial Cleaning: Drop variables from base_data
```{r}
str(base_data)
drop <- c("filename", "link", "batch", "selected_modules", "selected_window", "start", "end")
base = base_data[!(names(base_data) %in% drop)]
str(base)
```

B2. Initial Cleaning: Change "innonexpert" to "nonexpert"
```{r}
unique(base$type)
base$type <- as.factor(base$type)
# Rename by name: change "innonexpert" to "nonexpert"
levels(base$type)[levels(base$type)=="innonexpert"] <- "nonexpert"
unique(base$type)
```

C.1. Merge base + demo
```{r}
m1 <- merge(base, demo, by= "userId")
head(m1)

```

```

...
C.2.Join all_compare <- base + compare
```{r}
str(compare)
m2 <- merge(m1,compare, by= "recordid")
str(m2)
head(m2)
...

C.3. Access Module Look Up
```{r}
str(module)
...

C.4. Link Module_lookup to m2 in order to get spch_type for each clip
```{r}
drop3 <- c("Module", "Strapi.ID", "Title", "Cue.Strength", "No.of.Panels", "Rater.Study.Tasks")
module = module[,!(names(module) %in% drop3)]
str(module)
...

C.4. QA Data - lcross check values across m2 and m3

```{r}
chk = subset(m2, recordid == 44098642)
#str(chk)
...

C.5. Merge Module types with m2
```{r}
m3 <- merge(module, m2, by = "modul_panel")
head(m3)
...

C.6. Keep only Competence and Likeability-focused tasks
```{r}
m4 <- subset(m3, (Expected.Speech.Type == "Competence") | (Expected.Speech.Type == "Likeability"))
unique(m4$Expected.Speech.Type)
head(m4)
...

C.7. Change column name for spch_type and female = 1
```{r}
colnames(m4)[2] <- "spch_type"
str(m4)
colnames(m4)[10] <- "female"
...

C.8.Sort by userId, Module, window, recordid
```{r}
m4 <-m4[order(m4$userId, m4$modul_panel, m4>window_rank, m4$recordid),]
#head(m4,50)
...

C.9. Drop additional variables
```{r}
drop4 <- c("modul_panel", "Linking", "id", "window_rank", "ResponseId", "survey_type")
m5 = m4[,!(names(m4) %in% drop4)]
...

C.10. Move variables:
```{r}
install.packages("dplyr")
library(dplyr)

```

```

m5 <- m5 %>% relocate(userId, .before = spch_type)
m5 <- m5 %>% relocate(type, .before = recordid)
m5 <- m5 %>% relocate(female, .before = type)
str(m5)
nrow(m5)
'''

C.11. Rename spch_type levels: competence = C; Likeability = L // rename female levels: female = 1; male = 0
'''{r}
# Rename by name: change "innonexpert" to "nonexpert"
m5$spch_type <- as.factor(m5$spch_type)
levels(m5$spch_type)[levels(m5$spch_type)=="Competence"] <- "C"
levels(m5$spch_type)[levels(m5$spch_type)=="Likeability"] <- "L"
unique(m5$spch_type)

m5$female <- as.factor(m5$female)
unique(m5$female)
levels(m5$female)[levels(m5$female)=="Female"] <- "1"
levels(m5$female)[levels(m5$female)=="Male"] <- "0"
unique(m5$female)
write.csv(m5, "/cloud/project/all.compare.csv")
'''

C.12. Test - Remove faulty recordingids with faulty recordings per zero-values across acoustic features
'''{r}
library(dplyr)
nrow(m5)
faulty.ids <- as.list(read.csv("/cloud/project/remove.recordids.csv"))
length(faulty.ids$recordid)
m6 <- m5
omit.s <- nrow(m6)#5,065
m6 <- filter(m6, !(recordid %in% faulty.ids$recordid))
omit.f <- nrow(m6)#5,025
omit.f - omit.s #total number of rows omitted
'''

D. Create file of nonexpert speakers for further cleaning; identify number of speakers
'''{r}
nonexpert.compare <- subset(m6, type == "nonexpert")
nrow(nonexpert.compare)# nonexpert clips = 3,345
'''

D.2. File structure
'''{r}
ncol(nonexpert.compare)
#str(nonexpert.compare) #userId [,1]; female [,2]; type(ex/non) [,3]; recordid [,4]; spch_type [,5]; acoustic features
[,6:6,372]
'''

D4. Identify the number of speakers in the file
'''{r}
length(unique(nonexpert.compare$userId)) #155 speakers
'''

D.2. Identify balance of clips from female/male speakers
'''{r}
summary(nonexpert.compare$female) #Female(1) = 1,328; Male(0) = 1,987
'''

D.3. Total number of clips
'''{r}
nrow(nonexpert.compare) #n = 3,345

```

```

...
D.4. Balance of competence-focused to likability-focused clips
```{r}
summary(nonexpert.compare$spch_type) # Competence = 1,416; Lik = 1,929
...
D.5. Initial look for zero-values
```{r}
summary(nonexpert.compare$f1)
...
D.6a. Check for missing data
```{r}
#Check for missing data
sapply(nonexpert.compare, function(x) sum(is.na(x))) #05.13.2022 confirmed no missing values
...
D.6. Sample Lattice Graphs - histograms
```{r}
library(lattice)
histogram(~ f1 | factor(spch_type), data = nonexpert.compare)
...
D.8. Sample Lattice Graphs - QQ Plots
```{r}
qqmath(~ f1 | factor(spch_type), nonexpert.compare,
  f.value = ppoints(100), auto.key = TRUE,type = c("p", "g"), aspect = "xy")
#unique(nonexpert.compare$female)
...
D.9. Sample Lattice Graphs - BW Plots
```{r}
library(lattice)
library(gridExtra)

bw_theme <- trellis.par.get()
bw_theme$box.dot$spch <- ""
bw_theme$box.rectangle$col <- "black"
bw_theme$box.rectangle$lwd <- 2
bw_theme$box.rectangle$fill <- "grey90"
bw_theme$box.umbrella$ity <- 1
bw_theme$box.umbrella$col <- "black"
bw_theme$plot.symbol$col <- "grey40"
bw_theme$plot.symbol$pch <- "*"
bw_theme$plot.symbol$scex <- 2
bw_theme$strip.background$col <- "grey80"

lat1 <- bwplot(f1 ~ spch_type, data = nonexpert.compare, par.settings = bw_theme)
lat2 <- bwplot(f2 ~ spch_type, data = nonexpert.compare, par.settings = bw_theme)
lat3 <- bwplot(f3 ~ spch_type, data = nonexpert.compare, par.settings = bw_theme)
lat4 <- bwplot(f4 ~ spch_type, data = nonexpert.compare, par.settings = bw_theme)
lat5 <- bwplot(f5 ~ spch_type, data = nonexpert.compare, par.settings = bw_theme)
lat6 <- bwplot(f6 ~ spch_type, data = nonexpert.compare, par.settings = bw_theme)
lat7 <- bwplot(f7 ~ spch_type, data = nonexpert.compare, par.settings = bw_theme)
lat8 <- bwplot(f8 ~ spch_type, data = nonexpert.compare, par.settings = bw_theme)
lat9 <- bwplot(f9 ~ spch_type, data = nonexpert.compare, par.settings = bw_theme)
grid.arrange(lat1, lat2, lat3, lat4, ncol = 2)
grid.arrange(lat5, lat6, lat7, lat8, ncol = 2)
...

```

## R-Scripts - Data Partitioning

```

---
title: "DataPartitioning"
output: html_notebook
cr: 05-12-2022
mod: 05-13-2022
---
#https://rpubs.com/phamdinhkhanh/389752
#https://topepo.github.io/caret/variable-importance.html
#Inherit nonexpert.compare

```{r}
library("caret")
library("randomForest")
library(splitTools)
data.all <- nonexpert.compare
```

```{r}
str(data.all)
f <- data.all[data.all$female == "1", ]
length(unique(f$userId)) #n = 67
m <- data.all[data.all$female == "0", ]
length(unique(m$userId)) # n = 87
length(unique(data.all$userId))
```

```{r}
split.ids <- splitTools::partition(
  data.all$userId,
  p = c(train = 0.7, test = 0.3),
  type = "grouped"
)
train <- data.all[split.ids$train, ]
ncol(train)
drop8 <- c("userId")
train = train[!(names(train) %in% drop8)] #drop userId
train.labels <- train$spch_type
```

```{r}
nrow(train) #n = 2,413 audio clips
```

```{r}
table(train$spch_type) #C = 1,030, L = 1,383
```

```{r}
table(train$female) #female(1) = 1023; male(0) = 1,360
```

```{r}
table(train$female, train$spch_type)
```

```{r}
test <- data.all[split.ids$test, ]
test = test[!(names(test) %in% drop8)] #drop userId
test.labels <- test$spch_type
```

```{r}
nrow(test) #n = 972 audio clips

```

```
```\n```\n```\n{r}\ntable(test$spch_type) #C = 397 clips, L = 575 clips\n```\n```\n{r}\ntable(test$spch_type) #C = 397 clips, L = 575 clips\n```\n```\n{r}\ntable(test$female, test$spch_type)\n```\n```\n{r}\nstr(test)\n```\n\nDrop 'female' and 'type' from train and test sets\n```\n{r}\ndrop9 <- c("female", "type", "recordid")\ntrain = train[,!(names(train) %in% drop9)] #drop female, type, and recordid\ntest = test[,!(names(test) %in% drop9)] #drop female, type, and recordid\nhead(train)\nncol(test)\n```\n
```



## R-Scripts - Lasso

```

---
title: "Lasso"
output: html_notebook
cr: 05-15-2022
mod: 05-15-2022
---
A. Inherit and setup files
```{r}
library(caret)
library(glmnet)
library(pROC)
train <- read.csv("/Users/s.corrigan/Desktop/Dissertation/6.Dissertation_Data/DataMay16/train.csv")
train <- train[,-1] #removes a reference column at front of the df
test <- read.csv("/Users/s.corrigan/Desktop/Dissertation/6.Dissertation_Data/DataMay16/test.csv")
test <- test[,-1] #removes a reference column at front of the df

train.L1 <- train[,1:102]
test.L1 <- test[,1:102]
str(test[,1])
train.L1$spch_type <- as.numeric(train.L1$spch_type)-1 #prep outcome for application of glmnet
test.L1$spch_type <- as.numeric(test.L1$spch_type)-1
```

A2. Balance of classes
```{r}
table(train$spch_type) #"C" = 1030 (42.69%) "L" = 1383 (57.31%)
```

```{r}
table(test$spch_type) #N = 968 "C" = 397 (41.01%) "L" = 575 (59.40%)
```

B. Train L1 Logistic Regression
1. Train
```{r}
library(glmnet)
train.matrix <- model.matrix(spch_type ~., train.L1)[,-1]
lambdas <- 10^seq(8, -4, length = 250)

set.seed(25)
#L1Fit <- cv.glmnet(train.matrix,
  train.L1$spch_type,
  alpha = 1,
  lambda = lambdas,
  family = "binomial",
  probabilities = TRUE)
L1Fit
```

B2. Plot of MSE as a function of lambda
```{r}
dim(coef(L1Fit))
plot(L1Fit)
```

B3. Model
```{r}
L1Fit
```

B4. Model Results - best lasso value, i.e. minimizes training MSE

```

```

```{r}
lambda.lasso <- L1Fit$lambda.min
lambda.lasso
```

B5. Identify coefficients lasso drove to zero and those it kept
```{r}
set.seed(17)
predict(L1Fit, type = "coefficients", s = lambda.lasso) #Intercept, f2, f35, f56, f65
```

C. Lasso Performance
1. Performance on training set
```{r}
set.seed(26)
optlasso <- glmnet(train.matrix,
  train.L1$spch_type,
  alpha = 1,
  lambda = lambda.lasso,
  family = "binomial")
train.prob = predict(optlasso, train.matrix, type = "response")
train.class = train.prob > 0.5
train.class <- unlist(train.class)
train.class <- as.numeric(train.class)
train.L1$spch_type <- as.numeric(train.L1$spch_type)
mean(train.class == train.L1$spch_type) #0.5794
```

C2. Confusion Matrix - check to be sure variables are factors wth the same levels
```{r}
train.class <- as.factor(train.class)
levels(train.class) <- c("C", "L")
str(train.class)
```

```{r}
str(train$spch_type)
```

```{r}
confusionMatrix(train.class, train$spch_type)
```

C3. AUC
```{r}
library(pROC)
roc_score.train.L1 = roc(train$spch_type, train.prob , auc = TRUE, ci = TRUE) #AUC score
roc_score.train.L1
```

D1. Performance on test set
```{r}
test.matrix <- model.matrix(spch_type ~., test.L1)[-1]
test.prob <- predict(optlasso, s = lambda.lasso, newx = test.matrix, type = "response")
test.class <- (test.prob > 0.5)
str(test.class)
mean(test.class == test.L1$spch_type) #0.596
```

D2. Test Set Confusion Matrix
```{r}
test.CM = table(test.class, test.L1$spch_type)
test.CM
```

```

## D3. Error Metrics Function

```

```{r}
err_metric=function(CM)
{
  TN =CM[1,1]
  TP =CM[2,2]
  FP =CM[1,2]
  FN =CM[2,1]
  precision =(TP)/(TP+FP)
  recall_score =(FP)/(FP+TN)
  f1_score=2*((precision*recall_score)/(precision+recall_score))
  accuracy_model =(TP+TN)/(TP+TN+FP+FN)
  False_positive_rate =(FP)/(FP+TN)
  False_negative_rate =(FN)/(FN+TP)
  print(paste("Precision value of the model: ",round(precision,2)))
  print(paste("Accuracy of the model: ",round(accuracy_model,2)))
  print(paste("Recall value of the model: ",round(recall_score,2)))
  print(paste("False Positive rate of the model: ",round(False_positive_rate,2)))
  print(paste("False Negative rate of the model: ",round(False_negative_rate,2)))
  print(paste("f1 score of the model: ",round(f1_score,2)))
}
```

```

## D4. Performance Metrics - General

```

```{r}
err_metric(test.CM)
```

```

## D5. Performance Metrics - sensitivity / specificity directly

```

```{r}
test$spch_type <- as.factor(test$spch_type) #factor levels: "C" and "L"
test.class <- as.factor(test.class) #Change factor levels to "C" and "L" as well for confusionMatrix function
levels(test.class) <- c("C", "L")
```

```{r}
str(test.class)
unique(test.class)
```

```{r}
str(test$spch_type)
unique(test$spch_type)
```

```{r}
confusionMatrix(test.class, test$spch_type) #predicted value, expected value
```

```

## D6. Performance Metrics - AUC

```

```{r}
library(pROC)
roc_score.L1.test = roc(test.L1$spch_type, test.prob , auc = TRUE, ci = TRUE) #AUC score
roc_score.L1.test
```

```

## R-Scripts – Support Vector Classifier

```

---
title: "SupportVectorClassifier"
output:
  pdf_document: default
  html_notebook: default
  word_document: default
cr: 05-15-2022
mod: 05-15-2022
---
A1. Setup train and test data for support vector classifier
```{r}
train <- read.csv("/Users/s.corrigan/Desktop/Dissertation/6.Dissertation_Data/DataMay16/train.csv")
train <- train[,-1] #removes a reference column at front of the df
test <- read.csv("/Users/s.corrigan/Desktop/Dissertation/6.Dissertation_Data/DataMay16/test.csv")
test <- test[,-1] #removes a reference column at front of the df
train.data <- train
test.data <- test
```

A2. Scale and center train and test data
```{r}
train.data <- as.data.frame(scale(train.data[,2:101]))
train.data$spch_type <- train$spch_type
train.data <- train.data[,c(101, 1:100)]
test.data <- as.data.frame(scale(test.data[,2:101]))
test.data$spch_type <- test$spch_type
test.data <- test.data[,c(101, 1:100)]
#summary(train.data[,2]) # = 0
#sd(train.data[,2]) # = 1
```

B. Support vector classifier
```{r}
#Linear
library(caret)
train.Control <- trainControl(method="repeatedcv", number=10, repeats=3, classProbs = TRUE)
set.seed(254)
#svcFit <- train(train.data[,2:101], train.data[,1],
  #method = "svmLinear",
  #trControl = train.Control,
  #preProcess = c("center", "scale"),
  #tuneGrid = expand.grid(C = seq(from=0.5, to=6, by = 0.5)))
svcFit
```

C1. TRAIN DATA: Performance for SVC - Confusion Matrix
```{r}
###TRAIN###SVC
svc.classes.train <- predict(svcFit, newdata = train.data[,2:101])
confusionMatrix(data = svc.classes.train, train.data[,1])
```

C2. TRAIN DATA: Performance for SVC - AUC
```{r}
library(caret)
library(ROCR)
svc.probs.train <- predict.train(object = svcFit, newdata = train.data[,2:101], type='prob')[,1]
isPositiveClass.svc.train <- (train.data[,1] == 'C') # Define positive class
pred.svc.train <- prediction(svc.probs.train, isPositiveClass.svc.train)

```

```
perf.svc.train <- performance(pred.svc.train, 'tpr', 'fpr')

AUC.svc.train <- attributes(performance(pred.svc.train, 'auc'))$y.value[[1]] # area under curve for the svc
AUC.svc.train
```

D1. TEST DATA: Performance for SVC - Confusion Matrix
```{r}
svc.classes.test <- predict(svcFit, newdata = test.data[,2:101])
confusionMatrix(data = svc.classes.test, test.data[,1])
```

D2. TEST DATA: Performance for SVC - AUC
```{r}
#TEST SET#####TEST SET#####SVC
# prediction probabilities of test data classes
library(caret)
library(ROCR)
svc.probs.test <- predict.train(object = svcFit, newdata = test.data[,2:101], type='prob')[,1]
isPositiveClass.test <- test.data[,1] == 'C' # for a ROC curve there is a positive class (true match rate...) - defining
that class here
pred.svc.test <- prediction(svc.probs.test, isPositiveClass.test)
perf.svc.test <- performance(pred.svc.test, 'tpr', 'fpr')

AUC.svc.test <- attributes(performance(pred.svc.test, 'auc'))$y.value[[1]] # area under curve for the svc
AUC.svc.test
```
```

## R-Scripts - Support Vector Machine

```

---
title: "SupportVectorMachine_RadialKernel"
output: html_notebook
cr: 05-15-2022
mod: 05-19-2022
---
A. Setup train and test data for support vector machine
```{r}
train <- read.csv("/Users/s.corrigan/Desktop/Dissertation/6.Dissertation_Data/DataMay16/train.csv")
train <- train[,-1] #removes a reference column at front of the df
test <- read.csv("/Users/s.corrigan/Desktop/Dissertation/6.Dissertation_Data/DataMay16/test.csv")
test <- test[,-1] #removes a reference column at front of the df
train.data <- train
str(train.data)
test.data <- test
```

A2. Scale and center train and test data
```{r}
train.data <- as.data.frame(scale(train.data[,2:101]))
str(train.data)
train.data$spch_type <- train$spch_type
str(train.data)
train.data <- train.data[,c(101, 1:100)]
test.data <- as.data.frame(scale(test.data[,2:101]))
test.data$spch_type <- test$spch_type
test.data <- test.data[,c(101, 1:100)]
```

B. TRAIN: Support vector machine - radial basis
```{r}
bootControl <- trainControl(number = 200, classProbs = TRUE)

set.seed(234)
#svmFit <- train(train.data[,2:101], train.data[,1],
  # method = "svmRadial",
  # tuneLength = 5,
  # trControl = bootControl,
  # preProcess = c("center","scale"))
svmFit
```

C1. TRAIN DATA: Performance for SVM - Confusion Matrix
```{r}
svm.classes.train <- predict(svmFit, newdata = train.data[,2:101])
confusionMatrix(data = svm.classes.train, train.data[,1])
```

C2. TRAIN DATA: Performance for SVM - AUC
```{r}
svm.probs.train <- predict.train(object = svmFit, newdata = train.data[,2:101], type='prob')[,1]
isPositiveClass.svm.train <- train.data[,1] == 'C' # for a ROC curve there is a positive class (true match rate...) -
defining that class here
pred.svm.train <- prediction(svm.probs.train, isPositiveClass.svc.train)
perf.svm.train <- performance(pred.svm.train, 'tpr', 'fpr')
AUC.svm.train <- attributes(performance(pred.svm.train, 'auc'))$y.value[[1]] # area under curve for the svc

```

```
AUC.svm.train
...
D1. APPLY TEST - SVM
````{r}
svm.classes.test <- predict(svmFit, newdata = test.data[,2:101])
...
D2. TEST DATA: Performance for SVM - Confusion Matrix
````{r}
confusionMatrix(data = svm.classes.test, test.data[,1])
...
D3. TEST DATA: Performance for SVM - AUC
````{r}
#TEST SET ##### SVM
# prediction probabilities of test data classes
library(caret)
library(ROCR)
svm.probs.test <- predict.train(object = svmFit, newdata = test.data[,2:101], type='prob')[,1]
isPositiveClass.test <- test.data[,1] == 'C' # for a ROC curve there is a positive class (true match rate...) - defining
that class here
pred.svm.test <- prediction(svm.probs.test, isPositiveClass.test)
perf.svm.test <- performance(pred.svm.test, 'tpr', 'fpr')

AUC.svm.test <- attributes(performance(pred.svm.test, 'auc'))$y.value[[1]] # area under curve for the svc
AUC.svm.test
...

```

### Appendix C.1: R Scripts for Chapter 3

All scripts for Chapter 3 are located and available here:

<https://drive.google.com/drive/folders/1ilTnsrBkrBZ9HwP9GPcoJgT29MPKYz28>

### Appendix C.2: Estimated Facets for Competence-Focused Speech

Parameter	Facet	Estimate	Standard Error
Intelligence	Rating Scale	-0.094	0.031
Motivation	Rating Scale	0.297	0.031
Energy	Rating Scale	0.611	0.031
Overall Competence	Rating Scale	-0.202	0.032
Step 1	Step/Threshold ( $\tau$ )	-2.376	0.025
Step 2	Step/Threshold ( $\tau$ )	-0.269	0.021
Step 3	Step/Threshold ( $\tau$ )	2.645	0.033
Rater 1	Rater	1.637	0.032
Rater 2	Rater	-0.040	0.031
Rater 3	Rater	0.203	0.031
Rater 4	Rater	-0.887	0.032
Rater 5	Rater	-1.243	0.032
Rater 6	Rater	0.521	0.031
Rater 7	Rater	0.442	0.031
Rater 8	Rater	-0.633	0.083

### Appendix C.3: Estimated Facets for Likability-Focused Speech

Parameter	Facet	Estimate	Standard Error
Friendliness	Rating Scale	-0.012	0.029
Care	Rating Scale	0.482	0.029
Warmth	Rating Scale	0.259	0.029
Overall Likability	Rating Scale	-0.029	0.029
Step 1	Step/Threshold ( $\tau$ )	-2.007	0.023
Step 2	Step/Threshold ( $\tau$ )	-0.132	0.021
Step 3	Step/Threshold ( $\tau$ )	2.138	0.032
Rater 1	Rater	0.869	0.029
Rater 2	Rater	-0.770	0.029
Rater 3	Rater	0.530	0.028
Rater 4	Rater	-0.917	0.029
Rater 5	Rater	-1.178	0.029
Rater 6	Rater	0.253	0.028
Rater 7	Rater	1.309	0.029
Rater 8	Rater	-0.098	0.076



### References

- Ackerman, F., Malouf, R., & Blevins, J. P. (2016). Patterns and discriminability in language analysis. *Word structure*, 9(2), 132-155.  
<https://doi.org/10.3366/word.2016.0091>
- Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied psychological measurement*, 2(4), 581-594.  
<https://doi.org/10.1177/014662167800200413>
- Anikin, A. (2020). A moan of pleasure should be breathy: The effect of voice quality on the meaning of human nonverbal vocalizations. *Phonetica*, 77(5), 327-349.  
<https://doi.org/10.1159/000504855>
- Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6), 345-379.  
<https://doi.org/10.1007/s00530-010-0182-0>
- Attardo, S., Eisterhold, J., Hay, J., & Poggi, I. (2003). Multimodal markers of irony and sarcasm. *Humor*, 16, 243– 60.  
<http://dx.doi.org/10.1515/humr.2003.012>
- Baker, R. S., Gowda, S. M., Wixon, M., Kalka, J., Wagner, A. Z., Salvi, A., ... & Rossi, L. (2012). Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra. *International Educational Data Mining Society*.  
<https://files.eric.ed.gov/fulltext/ED537205.pdf>
- Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, 27(1), 3-23.  
<https://doi.org/10.2307/3315487>
- Barclay, P. (2016). Biological markets and the effects of partner choice on cooperation and friendship. *Current Opinion in Psychology*, 7, 33-38.  
<http://dx.doi.org/10.1016/j.copsyc.2015.07.012>
- Barry, W. J. (1981). Prosodic functions revisited again! *Phonetica*, 38(5-6), 320-340.  
<http://dx.doi.org/10.1159/000260036>
- Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798-1828. <https://doi.org/10.1109/tpami.2013.50>
- Bhaskaran, N., Nwogu, I., Frank, M. G., & Govindaraju, V. (2011, March). Lie to me: Deceit detection via online behavioral learning. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)* (pp. 24-29). IEEE.  
<https://doi.org/10.1109/fg.2011.5771407>
- Borràs-Comes, J., Roseano, P., Vanrell, M. del Mar, & Prieto, P. (2011). Perceiving uncertainty: Facial gestures, intonation, and lexical choice. In C. Kirchhof, Z. Malisz, & P. Wagner (Eds.), *Proceedings of the 2nd Conference on Gesture and Speech in Interaction [GESPIN 2011]*. Bielefeld University.
- Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144-152).  
<http://dx.doi.org/10.1145/130385.130401>
- Bousmalis, K., Mehu, M., & Pantic, M. (2009, September). Spotting agreement and

- disagreement: A survey of nonverbal audiovisual cues and tools. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops* (pp. 1-9). IEEE. <http://dx.doi.org/10.1109/ACII.2009.5349477>
- Bousmalis, K., Morency, L. P., & Pantic, M. (2011, March). Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)* (pp. 746-752). IEEE. <http://dx.doi.org/10.1109/FG.2011.5771341>
- Boutsen, F. (2003). Prosody: The music of language and speech. *The ASHA leader*, 8(4), 6-8. <https://doi.org/10.1044/leader.FTR1.08042003.6>
- Bor, A. (2020). Evolutionary leadership theory and economic voting: Warmth and competence impressions mediate the effect of economic perceptions on vote. *The Leadership Quarterly*, 31(2), 101295. <http://dx.doi.org/10.1016/j.leaqua.2019.05.002>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. <https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf>
- Brosch, T., Grandjean, D., Sander, D., & Scherer, K. R. (2008). Behold the voice of wrath: Cross-modal modulation of visual attention by anger prosody. *Cognition*, 106(3), 1497-1503. <http://dx.doi.org/10.1016/j.cognition.2007.05.011>
- Brosnan, S. F., Salwiczek, L., & Bshary, R. (2010). The interplay of cognition and cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1553), 2699-2710. <http://dx.doi.org/10.1098/rstb.2010.0154>
- Bull, P., & Connelly, G. (1985). Body movement and emphasis in speech. *Journal of Nonverbal Behavior*, 9(3), 169-187. <http://dx.doi.org/10.1007/BF01000738>
- Burkhardt, F., Schuller, B. W., Weiss, B., & Wengner, F. (2011). "Would you buy a car from me?"—on the likability of telephone voices. *Proceedings of the Interspeech, Florence, Italy (2011)*, pp. 1557-1560. <http://dx.doi.org/10.21437/Interspeech.2011-469>
- Canale, M., & Swain, M. (1981). A theoretical framework for communicative competence. In A.S. Palmer, P.J.M. Groot, G.A. Trosper (Eds.), *The construct validation of tests of communicative competence* (pp. 31-35). TESOL Publications,
- Cannistrà, M., Masci, C., Ieva, F., Agasisti, T., & Paganoni, A. (2020). Not the magic algorithm: Modelling and early-predicting students dropout through machine learning and multilevel approach. *MOX-Modelling and Scientific Computing, Department of Mathematics, Politecnico di Milano, Via Bonardi*. <https://www.mate.polimi.it/biblioteca/add/qmox/41-2020.pdf>
- Cavé, C., Guaitella, I., Bertrand, R., Santi, S., Harlay, F., & Espesser, R. (1996, October). About the relationship between eyebrow movements and F0 variations. In *Proceeding of Fourth International Conference on Spoken Language Processing (ICSLP'96)*, 4, 2157-2178. doi: 10.1109/ICSLP.1996.607235
- Cheang, H. S., & Pell, M. D. (2008). The sound of sarcasm. *Speech Communication*, 50(5), 366-381. <http://dx.doi.org/10.1016/j.specom.2007.11.003>
- Chen, X. L., Ita Levitan, S., Levine, M., Mandic, M., & Hirschberg, J. (2020). Acoustic-prosodic and lexical cues to deception and trust: deciphering how people detect lies. *Transactions of the Association for Computational Linguistics*, 8, 199-214. [http://dx.doi.org/10.1162/tacl\\_a\\_00311](http://dx.doi.org/10.1162/tacl_a_00311)
- Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4), 213. <https://doi.org/10.1037/h0026256>
- Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

- <http://dx.doi.org/10.4324/9780203771587>
- Collins, S. A. (2000). Men's voices and women's choices. *Animal Behaviour*, 60(6), 773-780.  
<https://doi.org/10.1006/anbe.2000.1523>
- Corrales-Astorgano, M., Escudero-Mancebo, D., & González-Ferreras, C. (2018). Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome. *Speech Communication*, 99, 90-100.  
<https://doi.org/10.1016/j.specom.2018.03.006>
- Cover, T. M. (2006). *Elements of information theory*. John Wiley & Sons.  
<https://doi.org/10.1002/047174882x>
- Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology*, 40, 61-149.  
[https://doi.org/10.1016/S0065-2601\(07\)00002-0](https://doi.org/10.1016/S0065-2601(07)00002-0)
- Cummins, N., Epps, J., & Kua, J. M. K. (2012). A comparison of classification paradigms for speaker likeability determination. In *Thirteenth Annual Conference of the International Speech Communication Association*.  
<http://dx.doi.org/10.21437/Interspeech.2012-93>
- Curhan, J. R., & Pentland, A. (2007). Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology*, 92(3), 802. doi:10.1037/0021-9010.92.3.802
- D'mello, S. K., & Graesser, A. (2010). Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction*, 20(2), 147-187.  
<https://doi.org/10.1007/s11257-010-9074-4>
- D'mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR)*, 47(3), 1-36.  
<https://doi.org/10.1145/2682899>
- Dal Palù, D., De Giorgi, C., Astolfi, A., Lerma, B., & Buiatti, E. (2014). SounBe, a toolkit for designers dealing with sound projects. In *DS 77: Proceedings of the DESIGN 2014 13th International Design Conference* (pp. 2011-2022).  
<https://www.designsociety.org/publication/35342/SOUNBE%2C+A+TOOLKIT+FOR+DESIGNERS+DEALING+WITH+SOUND+PROJECTS>
- Darwin, C. J. (1975). On the dynamic use of prosody in speech perception. In *Structure and process in speech perception* (pp. 178-194). Springer. [http://dx.doi.org/10.1007/978-3-642-81000-8\\_11](http://dx.doi.org/10.1007/978-3-642-81000-8_11)
- De Bot, K., Lowie, W., & Verspoor, M. (2007). A dynamic systems theory approach to second language acquisition. *Bilingualism: Language and Cognition*, 10(1), 7-21.  
<http://dx.doi.org/10.1017/S1366728906002732>
- Dessalles, J. L. (2014). 20.1 A fundamental and neglected issue. *The Social Origins of Language*, 19, 284. doi:10.1093/acprof:oso/9780199665327.001.0001
- Dunbar, R. I. M. (1998). *Grooming, gossip, and the evolution of language*. Harvard University Press.
- Duvvuru, S., & Erickson, M. (2013). The effect of change in spectral slope and formant frequencies on the perception of loudness. *Journal of Voice*, 27(6), 691-697.  
<https://doi.org/10.1016/j.jvoice.2013.05.004>

- Eckert, P., & Rickford, J. R. (Eds.). (2001). *Style and sociolinguistic variation*. Cambridge University Press. <http://dx.doi.org/10.1017/CBO9780511613258>
- Eckes, T. (2011). Introduction to many-facet Rasch measurement. Peter Lang. <https://doi.org/10.3726/978-3-653-04844-5>
- Eisenbruch, A. B., & Krasnow, M. M. (2022). Why warmth matters more than competence: A new evolutionary approach. *Perspectives on Psychological Science*. PMID: 35748187. doi: 10.1177/17456916211071087
- Enders, M. M., & Ward, P. I. (1985). Conflict and cooperation in the group feeding of the social spider *Stegodyphus mimosarum*. *Behaviour*, 94(1-2), 167-182. <https://doi.org/10.1163/156853985X00325>
- Engelhard Jr, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge. <https://doi.org/10.4324/9780203073636>
- Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1(1), 49-98. <https://doi.org/10.1515/9783110880021.57>
- Eyben, F. (2015). *Real-time speech and music classification by large audio feature space extraction*. Springer. <https://doi.org/10.1007/978-3-319-27299-3>
- Eyben, F., Batliner, A., & Schuller, B. (2010, April). Towards a standard set of acoustic features for the processing of emotion in speech. In *Proceedings of Meetings on Acoustics 159ASA* (Vol. 9, No. 1, p. 060006). Acoustical Society of America. <http://dx.doi.org/10.1121/1.4739483>
- Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013, October). Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia* (pp. 835-838). <http://dx.doi.org/10.1145/2502081.2502224>
- Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., & Truong, K. P. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2), 190-202. <http://dx.doi.org/10.1109/TAFFC.2015.2457417>
- Fant, G. (1960). *Acoustic theory of speech production*. Walter de Gruyter. <https://doi.org/10.1515/9783110873429>
- Feinberg, D. R., Jones, B. C., DeBruine, L. M., Moore, F. R., Smith, M. J. L., Cornwell, R. E., Tiddeman, B. P., Boothroyd, L. G., & Perrett, D. I. (2005). The voice and face of woman: One ornament that signals quality? *Evolution and Human Behavior*, 26(5), 398-408. <http://dx.doi.org/10.1016/j.evolhumbehav.2005.04.001>
- Fernandes, J., Teixeira, F., Guedes, V., Junior, A., & Teixeira, J. P. (2018). Harmonic to noise ratio measurement-selection of window and length. *Procedia Computer Science*, 138, 280-285. <https://doi.org/10.1016/j.procs.2018.10.040>
- Feyereisen, P., Van de Wiele, M., & Dubois, F. (1988). The meaning of gestures: What can be understood without speech? *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, 8(1), 3-25. <http://hdl.handle.net/2078.1/53419>
- Filippi, P. (2016). Emotional and interactional prosody across animal communication systems: A comparative approach to the emergence of language. *Frontiers in Psychology*, 7, 1393. <https://doi.org/10.3389/fpsyg.2016.01393>

- Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77-83. <https://doi.org/10.1016/j.tics.2006.11.005>
- Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science*, *27*(2), 67-73. <https://doi.org/10.1177/0963721417738825>
- Fleureau, J., Guillotel, P., & Huynh-Thu, Q. (2012). Physiological-based affect event detector for entertainment video applications. *IEEE Transactions on Affective Computing*, *3*(3), 379-385. <https://doi.org/10.1109/t-affc.2012.2>
- Formolo, D., & Bosse, T. (2017, July). Human vs. computer performance in voice-based recognition of interpersonal stance. In *International Conference on Human-Computer Interaction* (pp. 672-686). Springer.
- Formolo, D., & Bosse, T. (2018, October). Extracting interpersonal stance from vocal signals. In *Proceedings of the 4th International Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction* (pp. 19-25). [http://dx.doi.org/10.1007/978-3-319-58071-5\\_51](http://dx.doi.org/10.1007/978-3-319-58071-5_51)
- Freedman, N., & Hoffman, S. P. (1967). Kinetic behavior in altered clinical states: Approach to objective analysis of motor behavior during clinical interviews. *Perceptual and Motor Skills*, *24*(2), 527-539. <http://dx.doi.org/10.2466/pms.1967.24.2.527>
- Friedman, J., Hastie, T., Tibshirani, R., & Narasimhan, B. (2021). Package 'glmnet'. *CRAN R Repository*. <https://cran.r-project.org/web/packages/glmnet/index.html>
- Gallardo, L. F., & Weiss, B. (2017). Perceived interpersonal speaker attributes and their acoustic features. In *Phonetik und Phonologie im Deutschsprachigen Raum (PundP'13)* (pp. 61-64).
- Garfinkel, H. (1967). *Studies in ethnomethodology*. Englewood Cliffs.
- Gee, M. G., Tomlins, P., Calver, A., Darling, R. H., & Rides, M. (2005). A new friction measurement system for the frictional component of touch. *Wear*, *259*(7-12), 1437-1442.
- Gigilashvili, D., Shi, W., Wang, Z., Pedersen, M., Hardeberg, J. Y., & Rushmeier, H. (2021). The role of subsurface scattering in glossiness perception. *ACM Transactions on Applied Perception (TAP)*, *18*(3), 1-26. <https://doi.org/10.1016/j.wear.2005.02.053>
- Geeslin, K. L., Gudmestad, A., Kanwit, M., Linford, B., Long, A. Y., Schmidt, L., & Solon, M. (2018). Sociolinguistic competence and the acquisition of speaking. *Speaking in a Second Language*, 1-25. <https://doi.org/10.1075/aals.17.01gee>
- Gigilashvili, D., Thomas, J. B., Pedersen, M., & Hardeberg, J. Y. (2019, October). Perceived glossiness: Beyond surface properties. In *Color and imaging conference* (Vol. 2019, No. 1, pp. 37-42). Society for Imaging Science and Technology. <https://doi.org/10.2352/issn.2169-2629.2019.27.8>
- Gilmore, D. C., Stevens, C. K., Harrell-Cook, G., & Ferris, G. R. (1999). Impression management tactics. In *The Employment Interview Handbook* (pp. 321-336). Chapter doi: <https://dx.doi.org/10.4135/9781452205519.n18>
- Gobl, C., & Chasaide, A. N. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, *40*(1-2), 189-212. [https://doi.org/10.1016/S0167-6393\(02\)00082-1](https://doi.org/10.1016/S0167-6393(02)00082-1)
- Goffman, E. (1967). *Interaction ritual: Essays on face-to-face interaction*. Routledge.

- Gonzalez, S., & Anguera, X. (2013, May). Perceptually inspired features for speaker likability classification. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8490-8494). IEEE. <http://dx.doi.org/10.1109/ICASSP.2013.6639322>
- Gupta, R., Lee, C. C., Bone, D., Rozga, A., Lee, S., & Narayanan, S. (2012). Acoustical analysis of engagement behavior in children. In *Third Workshop on Child, Computer and Interaction*. <https://doi.org/10.1145/1640377>
- Gustafsson, J. E. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, 33(2), 205-233. <https://doi.org/10.1111/j.2044-8317.1980.tb00609.x>
- Hack, T., Goodwin, S. A., Fiske, S. T. (2013). Warmth trumps competence in evaluations of both ingroup and outgroup. *International Journal of Science, Commerce and Humanities*, 1(6), 99-105.
- Hadar, U., Steiner, T. J., & Rose, F. C. (1984). The relationship between head movements and speech dysfluencies. *Language and Speech*, 27(4), 333-342. <http://dx.doi.org/10.1177/002383098402700404>
- Hadar, U., Steiner, T. J., Grant, E. C., & Rose, F. C. (1983). Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, 2(1-2), 35-46. [http://dx.doi.org/10.1016/0167-9457\(83\)90004-0](http://dx.doi.org/10.1016/0167-9457(83)90004-0)
- Hadar, U., Steiner, T. J., Grant, E. C., & Rose, F. C. (1984). The timing of shifts of head postures during conversation. *Human Movement Science*, 3(3), 237-245. [http://dx.doi.org/10.1016/0167-9457\(84\)90018-6](http://dx.doi.org/10.1016/0167-9457(84)90018-6)
- Haiman, J. (1998). Talk is cheap: Sarcasm, alienation, and the evolution of language. *Oxford University Press on Demand*. doi: 10.1017/S0047404500211032
- Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian Journal of Internal Medicine*, 4(2), 627.
- Hammarberg, B., Fritzell, B., Gaufin, J., Sundberg, J., & Wedin, L. (1980). Perceptual and acoustic correlates of abnormal voice qualities. *Acta Oto-Laryngologica*, 90(1-6), 441-451. <https://doi.org/10.3109/00016488009131746>
- Han, H., Guo, X., & Yu, H. (2016, August). Variable selection using mean decrease accuracy and mean decrease gini based on random forest. In *2016 7th IEEE International Conference On Software Engineering And Service Science (icsess)* (pp. 219-224). IEEE. 10.1109/ICSESS.2016.7883053
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer. <http://dx.doi.org/10.1007/978-0-387-84858-7>
- Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, 78(3), B53-B64. [https://doi.org/10.1016/s0010-0277\(00\)00132-3](https://doi.org/10.1016/s0010-0277(00)00132-3)
- Hillenbrand, J., & Houde, R. A. (1996). Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *Journal of Speech, Language, and Hearing Research*, 39(2), 311-321. doi: 10.1044/jshr.3902.311
- Heritage, J. (2013). Action formation and its epistemic (and other) backgrounds. *Discourse Studies*, 15(5), 551-578. <http://dx.doi.org/10.1177/1461445613501449>
- Hermansky, H., & Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4), 578-589. <http://dx.doi.org/10.1109/89.326616>

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer. <http://dx.doi.org/10.1007/978-1-0716-1418-1>
- Janitza, S., Tutz, G., & Boulesteix, A. L. (2016). Random forest for ordinal responses: Prediction and variable selection. *Computational Statistics & Data Analysis*, 96, 57-73. <http://dx.doi.org/10.1016/j.csda.2015.10.005>
- Jesse, A., & Johnson, E. K. (2012). Prosodic temporal alignment of co-speech gestures to speech facilitates referent resolution. *Journal of Experimental Psychology: Human Perception and Performance*, 38(6), 1567. <http://dx.doi.org/10.1037/a0027921>
- Ji, J., Hu, L., Liu, B., & Li, Y. (2020). Identifying and assessing the impact of key neighborhood-level determinants on geographic variation in stroke: a machine learning and multilevel modeling approach. *BMC Public Health*, 20(1), 1-12. DOI:10.1186/s12889-020-09766-3
- Kendon, A. (1994). Do gestures communicate? A review. *Research on Language and Social Interaction*, 27(3), 175-200. doi:10.1207/s15327973rlsi2703\_2
- Kennedy, C. J., Bacon, G., Sahn, A., & von Vacano, C. (2020). Constructing interval variables via faceted Rasch measurement and multitask deep learning: A hate speech application. *arXiv preprint arXiv:2009.10277*. <https://arxiv.org/pdf/2009.10277.pdf>
- Kent, R. D., Kent, R. A., & Read, C. (2002). *The acoustic analysis of speech*. Singular.
- Kiesling, S. F. (2009). Style as stance. In *Stance: Sociolinguistic perspectives* (pp. 171-194). Oxford University Press.
- Kimble, C. E., & Seidel, S. D. (1991). Vocal signs of confidence. *Journal of Nonverbal Behavior*, 15(2), 99-105. <http://dx.doi.org/10.1007/BF00998265>
- Klofstad, C. A., Anderson, R. C., & Peters, S. (2012). Sounds like a winner: Voice pitch influences perception of leadership capacity in both men and women. *Proceedings of the Royal Society B: Biological Sciences*, 279(1738), 2698-2704. <https://doi.org/10.1098/rspb.2012.0311>
- Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5), 1-26. <http://dx.doi.org/10.18637/jss.v028.i05>
- Leary, T. (1957). *Interpersonal diagnosis of personality*. Ronald'Press Co.
- Lee, C. M., Narayanan, S. S., & Pieraccini, R. (2002, September). Combining acoustic and language information for emotion recognition. In *INTERSPEECH*. <https://sail.usc.edu/publications/files/leeinterspeech2002.pdf>
- Lehiste, I., & Lass, N. J. (1976). Suprasegmental features of speech. *Contemporary issues in experimental phonetics*, 225, 239. <https://doi.org/10.1016/b978-0-12-437150-7.50013-0>
- Leloup, F. B., Obein, G., Pointer, M. R., & Hanselaer, P. (2014). Toward the soft metrology of surface gloss: A review. *Color Research & Application*, 39(6), 559-570. <https://doi.org/10.1002/col.21846>
- Levitan, S. I., Maredia, A., & Hirschberg, J. (2018, June). Linguistic cues to deception and perceived deception in interview dialogues. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 1941-1950). <http://dx.doi.org/10.18653/v1/N18-1176>
- Linacre, J. M. (1989). *Many-faceted Rasch measurement* (Doctoral dissertation, The University of Chicago). <https://www.proquest.com/>
- Linacre, J. M. (2010). Predicting responses from Rasch measures. *Journal of Applied Measurement*, 11(1), 1.



- <https://winsteps.com/a/Linacre-Predicting.pdf>
- Ludwig, S., Van Laer, T., De Ruyter, K., & Friedman, M. (2016). Untangling a web of lies: Exploring automated detection of deception in computer-mediated communication. *Journal of Management Information Systems*, 33(2), 511-541. <https://doi.org/10.1080/07421222.2016.1205927>
- Lukowicz, P., Pentland, S., & Ferscha, A. (2011). From context awareness to socially aware computing. *IEEE Pervasive Computing*, 11(1), 32-41. doi: 10.1109/MPRV.2011.82
- MacIntyre, P. D., & Ayers-Glassey, S. (2020). 7. Competence appraisals: Dynamic judgements of communication competence in real time. In *Usage-based dynamics in second language development* (pp. 155-175). Multilingual Matters. <https://doi.org/10.21832/9781788925259-010>
- Mari, L., & Wilson, M. (2014). An introduction to the Rasch measurement approach for metrologists. *Measurement*, 51, 315-327. <https://doi.org/10.1016/j.measurement.2014.02.014>
- Mari, L., Wilson, M., & Maul, A. (2021). *Measurement across the sciences*. Springer. <https://doi.org/10.1007/978-3-030-65558-7>
- Marler, P. (1967). Animal Communication Signals: We are beginning to understand how the structure of animal signals relates to the function they serve. *Science*, 157(3790), 769-774. <https://doi.org/10.1126/science.157.3790.769>
- McDonald, S. (1992). Differential pragmatic language loss after closed head injury: Ability to comprehend conversational implicature. *Applied Psycholinguistics*, 13(3), 295-312. doi: <https://doi.org/10.1017/S0142716400005658>
- McNeill, D. (Ed.). (2000). *Language and gesture* (Vol. 2). Cambridge University Press. <http://dx.doi.org/10.1017/CBO9780511620850>
- Nembrini, S., König, I. R., & Wright, M. N. (2018). The revival of the Gini importance? *Bioinformatics*, 34(21), 3711-3718. doi: 10.1093/bioinformatics/bty373
- Nesse, R. M. (2007). Runaway social selection for displays of partner value and altruism. *Biological Theory*, 2(2), 143-155. [http://dx.doi.org/10.1007/978-1-4020-6287-2\\_10](http://dx.doi.org/10.1007/978-1-4020-6287-2_10)
- Noë, R. (2006). Cooperation experiments: Coordination through communication versus acting apart together. *Animal Behaviour*, 71(1), 1-18. <https://doi.org/10.1016/j.anbehav.2005.03.037>
- Ochs, E. (1992). 14 Indexing gender. *Rethinking context: Language as an interactive phenomenon*, 11, 335.
- Ochs, E. (1993). Constructing social identity: A language socialization perspective. *Research on Language and Social Interaction*, 26(3), 287-306. [https://doi.org/10.1207/s15327973rlsi2603\\_3](https://doi.org/10.1207/s15327973rlsi2603_3)
- Oguchi, T., & Kikuchi, H. (1997). Voice and interpersonal attraction. *Japanese Psychological Research*, 39(1), 56-61. doi:10.1111/1468-5884.00037
- Palmer, C., & Hutchins, S. (2006). What is musical prosody?. *Psychology of Learning and Motivation*, 46, 245-278. [http://dx.doi.org/10.1016/S0079-7421\(06\)46007-2](http://dx.doi.org/10.1016/S0079-7421(06)46007-2)
- Parhankangas, A., & Ehrlich, M. (2014). How entrepreneurs seduce business angels: An impression management approach. *Journal of Business Venturing*, 29(4), 543-564. <https://doi.org/10.1016/j.jbusvent.2013.08.001>



- Partan, S., & Marler, P. (1999). Communication goes multimodal. *Science*, 283(5406), 1272-1273. <http://dx.doi.org/10.1126/science.283.5406.1272>
- Partan, S. R., & Marler, P. (2005). Issues in the classification of multimodal communication signals. *The American Naturalist*, 166(2), 231-245. <https://doi.org/10.1086/431246>
- Paulmann, S., Titone, D., & Pell, M. D. (2012). How emotional prosody guides your way: Evidence from eye movements. *Speech Communication*, 54(1), 92-107. <http://dx.doi.org/10.1016/j.specom.2011.07.004>
- Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, 46(4), 1023-1031. <https://doi.org/10.3758/s13428-013-0434-y>
- Pentland, A. (2010). *Honest signals: How they shape our world*. MIT press. <https://doi.org/10.1016/j.jbusvent.2013.08.001>
- Pon-Barry, H., & Shieber, S. M. (2011). Recognizing uncertainty in speech. *EURASIP Journal on Advances in Signal Processing*, 2011, 1-11. <https://doi.org/10.1155/2011/251753>
- Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37, 98-125. <https://doi.org/10.1016/j.inffus.2017.02.003>
- Priva, U. C. (2015). Informativity affects consonant duration and deletion rates. *Laboratory phonology*, 6(2), 243-278. <https://doi.org/10.1515/lp-2015-0008>
- Rabinov, C. R., Kreiman, J., Gerratt, B. R., & Bielamowicz, S. (1995). Comparing reliability of perceptual ratings of roughness and acoustic measures of jitter. *Journal of Speech, Language, and Hearing Research*, 38(1), 26-32. doi: 10.1044/jshr.3801.26
- Ranganath, R., Jurafsky, D., & McFarland, D. A. (2013). Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. *Computer Speech & Language*, 27(1), 89-115. <https://doi.org/10.1016/j.csl.2012.01.005>
- Regan, V. (2010). Sociolinguistic competence, variation patterns and identity construction in L2 and multilingual speakers. *Eurosla Yearbook*, 10(1), 21-37.
- Rigoulot, S., & Pell, M. D. (2012). Seeing emotion with your ears: Emotional prosody implicitly guides visual attention to faces. *PloS One*, 7(1), Article e30740. <http://dx.doi.org/10.1371/journal.pone.0030740>
- Scherer, K. R., London, H., & Wolf, J. J. (1973). The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality*, 7(1), 31-44. [https://doi.org/10.1016/0092-6566\(73\)90030-5](https://doi.org/10.1016/0092-6566(73)90030-5)
- Schegloff, E. A. (1997). Third turn repair. *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, 31-40. <http://dx.doi.org/10.1075/cilt.128.05sch>
- Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., van Son, R., Wenginger, F., Eyben, F., Bocklet, T., Mohammadi, G., & Weiss, B. (2012). The INTERSPEECH 2012 speaker trait challenge. In *INTERSPEECH 2012, Portland, OR, USA*. <https://doi.org/10.21437/interspeech.2012-86>
- Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Wenginger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., & Kim, S. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH*

- 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France. <https://doi.org/10.21437/interspeech.2013-56>
- Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., van Son, R., Wenginger, F., Eyben, F., Bocklet, T., Mohammadi, G., & Weiss, B. (2015). A survey on perceived speaker traits: Personality, likability, pathology, and the first challenge. *Computer Speech & Language*, 29(1), 100-131. <https://doi.org/10.1016/j.csl.2014.08.003>
- Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCr: Visualizing classifier performance in r. *Bioinformatics*, 21(20), 3940-3941. <http://dx.doi.org/10.1093/bioinformatics/bti623>
- Staples, E. J. (2000). The zNose, a new electronic nose using acoustic technology. *Journal of the Acoustical Society of America*, 108(5), 2495. <https://doi.org/10.1121/1.4743211>
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677-680. <https://doi:10.1037/hoo56886>
- Sutherland, W. J. (2005). The best solution. *Nature*, 435(7042), 569-569. <https://doi.org/10.1038/435569a>
- Tigue, C. C., Borak, D. J., O'Connor, J. J., Schandl, C., & Feinberg, D. R. (2012). Voice pitch influences voting behavior. *Evolution and Human Behavior*, 33(3), 210-216. <https://doi.org/10.1016/j.evolhumbehav.2011.09.004>
- Vallejo, M., De La Espriella, C., Gómez-Santamaría, J., Ramírez-Barrera, A. F., & Delgado-Trejos, E. (2019). Soft metrology based on machine learning: a review. *Measurement Science and Technology*, 31(3), 032001. <https://doi.org/10.1088/1361-6501/ab4b39>
- Verspoor, M. (2013). Dynamic systems theory as a comprehensive theory of second language development. *Contemporary Approaches to Second Language Acquisition*, 9, 199. doi:10.1075/aals.9.13ch10
- Verspoor, M. (2017). Complex dynamic systems theory and L2 pedagogy. *Complexity theory and Language Development: In celebration of Diane Larsen-Freeman*, 143-62. <https://doi.org/10.1075/llt.48.08ver>
- Vinciarelli, A., Pantic, M., Bourlard, H., & Pentland, A. (2008, October). Social signal processing: State-of-the-art and future perspectives of an emerging domain. In *Proceedings of the 16th ACM international conference on Multimedia* (pp. 1061-1070). <https://doi.org/10.1145/1459359.1459573>
- Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and vision computing*, 27(12), 1743-1759. <http://dx.doi.org/10.1016/j.imavis.2008.11.007>
- Vinciarelli, A., & Esposito, A. (2018). Multimodal analysis of social signals. In *The handbook of multimodal-multisensor interfaces: Signal processing, architectures, and detection of emotion and cognition* (Vol. 2) (pp. 203-226). <https://doi.org/10.1145/3107990.3107999>
- Wagner, J., Andre, E., Lingenfelser, F., & Kim, J. (2011). Exploring fusion methods for multimodal emotion recognition with missing data. *IEEE Transactions on Affective Computing*, 2(4), 206-218. <https://doi.org/10.1109/t-affc.2011.12>
- Wickler, W. (1978). A special constraint on the evolution of composite signals. *Zeitschrift für Tierpsychologie*, 48(4), 345-348. <https://doi.org/10.1111/j.1439-0310.1978.tb00265.x>

- Wiemann, J. M. (1977). Explication and test of a model of communicative competence. *Human Communication Research*, 3(3), 195-213. <https://doi.org/10.1111/j.1468-2958.1977.tb00518.x>
- Wildgruber, D., Ackermann, H., Kreifelts, B., & Ethofer, T. (2006). Cerebral processing of linguistic and emotional prosody: fMRI studies. *Progress in brain research*, 156, 249-268. [https://doi.org/10.1016/s0079-6123\(06\)56013-3](https://doi.org/10.1016/s0079-6123(06)56013-3)
- Wollum, E. (2019). The uptalk downgrade: Comparing age-and gender-based perceptions of uptalk in four highly skilled professions. [Master's thesis, Victoria University of Wellington]. VUW Research Archive.
- Wright, B., & Linacre, J. M. (1994). *Reasonable mean-square fit values*. Rasch. <https://www.rasch.org/rmt/rmt83b.htm>
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA press. <https://research.acer.edu.au/measurement/2/>
- Yumoto, E., Gould, W. J., & Baer, T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *The Journal of the Acoustical Society of America*, 71(6), 1544-1550. <https://doi.org/10.1121/1.387808>