

**Employing Machine Learning Techniques in Conjunction with Linguistic Scoring
Mechanism to Formalize the Uncertainty Carried by Scoring Numeric Processes Driven by
Probabilistic Methods**

By

Javier Livio

A dissertation submitted to the Graduate Faculty of

Auburn University

In partial fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Auburn, Alabama

December 12, 2020

Keywords: Speech Intelligibility; Direct Magnitude Estimation (DME); Selectable Scoring;
Machine Learning, Approximate Reasoning, Expert System, Fuzzy set Theory, Possibility
Theory, Uncertainty, Generalized Information Theory, Coffee Roasting, Coffee Judging

Copyright 2020 by Javier Livio

Approved by

Dr. Cheryl D. Seals, Chair, Professor of Computer Science and Software Engineering
Dr. Richard Chapman, Co-chair, Associate Professor of Computer Science and Software
Engineering

Dr. Marisha Speights Atkins, Assistant Professor of Communication Disorders

Dr. David Umphress, Professor of Cybersecurity and Information Assurance

Dr. Noah Dormady, Associate Professor, John Glenn College of Public Affairs, The Ohio State
University

Dr. Wilfredo Flores, Professor of Computer Science and Electrical Engineering, UNITEC

Abstract

Essay 1

This chapter aims to provide a model for quantifying judgment in order to set a foundation for human knowledge representation and reasoning while making decisions in the presence of vagueness, imprecision, and uncertainty. Instead of seeking exact values through deterministic systems, based on computer science mature algorithms and data-structures, this research empowers software engineering to handle the uncertainty, which pertains to non-deterministic systems, bridging their study, their modeling with techniques such as machine learning. The theoretical framework of this research is grounded in a group of theories, i.e. possibility theory and the uncertainty-based information theory as they guide us in the quantification of uncertainty and proposed mature techniques to study and represent imprecision. Probability theory has a wider application to uncertainty problems, the same is true for possibility theory and the general theory of uncertainty-based information due to imprecise or incomplete information (i.e. studied by fuzzy set theory). This research has gained motivation from the fact that an object may more or less become the representation of the category in which one attempts to place it even when the states of its entities are not well known (i.e., deriving the final grade of high-quality coffee beans from their individual attributes such as Fragrance, Aroma, etc., hypothetically impacted by how was roasted and, measuring children's speech intelligibility as the child utters words. Then, predicting these two systems outputs using machine learning models, trained and validated with collected parameters from non-deterministic systems featuring controlled variability). Two highly complex systems whose behavior is not well understood have been given an approximated solution (hypothetically with less error), not only by modeling themselves but, by modeling their uncertainty as well by utilizing both linguistic scoring techniques and machine learning. The aim

is to integrate the naturalistic mechanisms of human decision making with machine learning capabilities as a useful framework with the hypothetical capacity of supporting several domains of the human decision-making process.

Essay 2

This chapter presents a novel approach from the perspective of computational intelligence by applying fuzzy logic, machine learning. We proposed the unifications of roasting and cupping by combining the expert's knowledge from each of these domains, including a design of a Fuzzy Controller for modeling the roasting of raw coffee beans and their impact in the final quality grade given to high-quality coffees known as specialty coffees. This device's approach, theoretically, will roast coffee beans by looking at their final quality as an optimization problem to be solved, not as the coffee industry standard protocol indicates.

Essay 3

Speech intelligibility estimation lacks a standardized measurement system. This has produced an important clinical need for evaluative tools to functionally assess intelligibility in children, and to identify estimation approaches for conceptualizing the problem beyond subjective intuition. Young children, in particular, present a clinical challenge in speech intelligibility estimation. As speech becomes less intelligible, unfamiliar speech patterns become more difficult to assess even amongst highly-trained clinicians. As large data sets have become available, more advanced methods based on deep learning have yielded more intelligent solutions (beyond regression models) for solving complex computational problems. However further research is needed, specifically in the pediatric population, to develop a useful application of deep learning models for automatic speech intelligibility detection that captures both the abnormal speech variation and

subjective ratings of assessors of speech intelligibility. This research aims to fill part of this gap. We introduce a novel alternative supporting intelligibility assessment methodology for young children based on linguistic, rather than numeric scoring mechanisms. We also introduce a software tool to estimate speech intelligibility using the long-established Direct Magnitude Estimation (DME) approach. The assessment methodology is tested against the established approach with a randomized controlled trial (experiment). The experiment compares the linguistic scoring system against the numeric scoring system employed by inexperienced and experienced listeners. Econometric analysis of the experiment data, using linear regression estimation techniques, identifies important implications for the use of linguistic terminology for scoring and evaluating speech intelligibility. We find more variability in the evaluations of the group of trained clinicians in all scoring techniques i.e. (linguist, numeric, and both) versus the group of inexperienced listeners. The research results suggest the use of linguistic scoring mechanisms results in decreased variability (i.e., greater precision) in intelligibility assessment than the longstanding numeric scoring systems with both experimental groups. The implications of this methodology and these findings have great potential well beyond speech intelligibility, informing subjective assessments using numeric scoring systems. Furthermore, a machine learning model was developed employing a freshly captured dataset. DME metrics were gathered from trained clinicians presented only with a linguistic scoring option because this research suggests this methodology decreases variability.

Acknowledgments

I have been blessed with a well-equipped and highly creative committee.

I would like to thank my committee chair, Professor Cheryl Seals, for her support and guidance through each stage of the process. Without her patience and unwavering support, I would not have made it. For this, I'm profoundly grateful.

Dr Marisha Speights, she walked with me through the lands of speech pathology. For this, I am extremely grateful.

Dr Noah Dormady, was instrumental in defining the path of my research, he formed in me love for advanced statistics. For this, I am extremely grateful.

Dr. Richard Chapman. His guidance during difficult times and prompt support, his knowledge of various domains provided clear guide into my research.

Dr. David Umphress. He welcomes me into the Auburn family, he supported my research and sponsored my participation at the World Conference of Computational Intelligence in Rio de Janeiro, Brazil. His work ethic and passion for science have taken me to a higher level in my research. Thank you for challenging me.

Dr. Wilfredo Flores for inspiring my interest in the development of innovative technologies from the perspective of Fuzzy Logic and the coffee industry.

My beautiful wife, Ines Livio for believing in me and offering unconditional support throughout my whole career. We met at high school, since then, she has taught me passion, persistence and most of all to take responsibility of my decisions. For this, I am deeply grateful.

My kids, Sebastian, Karen and her husband. They inspired and support me. I am grateful for having them in my life.

To my dearest sister Margaret, she taught me how to read, she sparked in me the love for science.

Purpose and Problem Statement

Through a series of field observations, it was possible to develop a hypothesis about potential bias and variability in the numeric processes of scoring in some contexts (e.g., modeling the relationship between roasting and grading high-quality coffees beans, the measurement of child speech intelligibility through Direct Magnitude Estimation or DME).

To formalize human impression, there are two possible scenarios (whether or not is a bias or error) that will be tested in this research and a question to explore:

1. The existent methodologies (i.e., use of numeric scoring mechanism) inflates the variance/variability in the evaluative processes.
2. The existing methodologies introduces bias (e.g., error in a consistent direction).
3. Systems that support good decisions making, when designed from domain expert's knowledge, they should carry the less possible error.
4. Heuristics are quite useful, but sometimes they lead to severe and systematic errors.
5. The existing methodologies seem not to encourage human heuristics to potentially take less work/effort to report linguistic versus numeric values.
6. Solving complex computational problems by automated assessment and AI.

In this dissertation, we propose two separate studies to formally test the following overarching research questions:

Research Questions

1. Can Machine Learning techniques in conjunction with a linguistic scoring mechanism be utilized to address these problems of how decision-makers evaluate in numeric processes?
2. Can Machine Learning techniques be utilized to reduce bias and/or error by replacing numeric processes with linguistic ones that have the capacity to be more naturalistic and less stressful for human decision-makers?

Significance

Even if this research provides little evidence that there is a bias rather than error, this could extrapolate affecting the scope of several domains e.g. medical evaluations, prescriptions, control, planning, banking, trading, predictions, decision making, scheduling, diagnosis, etc.

In addition, if these experiments support evidence that it takes the subject less work/effort to report linguistic versus numeric values, the implication could lead to the re-design of data gathering in general.

Moreover, if machine learning techniques can offer a modest improvement, this could indicate, essentially a vital contribution of this research program. Similar industrial and societal applications include naturalistic language processing, sentiment analysis, facial recognition, accurate medical diagnoses of various illnesses [1], [2]. Furthermore, in decision-making, investments and, risk analysis; a broad spectrum of software applications meant to help investors. Would this help them make less erroneous investment decisions, result in more stable retirements, lower pension costs, more effective illness diagnosis (i.e., accurate early detections), an increase

on space exploration mission's success by having more efficient anomaly detection mechanisms in place, etc.?

Table of Contents

Abstract	ii
Acknowledgements	v
Purpose and Problem Statement	vi
Research Questions	vii
Significance.....	vii
List of Tables	xv
List of Figures	xvii
List of Abbreviations	xxii
Essay 1: Designing Machine Learning Models Leveraging Linguistic Scoring Uncertainty	
(De-erroring)	1
Introduction.....	1
Beyond the Organized Simplicity Perspective	2
Bias and Error	7
De-biasing	7
De-erroring	8
Recent Applications of Machine Learning Supporting Human	
Decision-Making	10
Supporting Theories in the Literature	11

Possibility Theory	11
Uncertainty-Based Information Theory	13
Fuzzy Logic	14
Deep Learning	16
Theoretical Problems to Address	17
Industry Problem: Coffee Bean Roasting and its Impact on their Final Quality Grade	18
Social Problem: Children Speech Intelligibly	19
Theoretical Approach to Solve Proposed Problems	20
Key Highlights of this Research	25
About Coffee Bean Roasting and its Impact on their Final Quality Grade	25
About Children Speech Intelligibly	26
Summary	27
Essay 2: Computational Intelligence: Fuzzy Logic and Machine Learning Applied to Modelling	
Raw Beans Roasting and Judging (Cupping) of Specialty Coffees.....	28
Introduction.....	28
Motivation and Study Problem	28
Purpose and Research Approach	28
The Goals of the Study are to Support the Following Research Contribution	29

Organization of this Work	30
Research Questions	30
Research Hypothesis	30
Limitations	31
Key Terms	31
Literature Review	32
Experimental Methods / Research Landscape	33
Hypothesized Functional Form of Cupping Results Driven by Roast Profile and Bean Parameters	34
Roasting Specialty Coffee Beans	36
Inferring Specialty Coffee Bean’s Quality Grading	37
Roaster Fuzzy Controller	40
Incorporating Machine Learning (Random Forest Regressor)	43
Discussion of Preliminary Results	44
Linear Regression Models, Looking for linearity in the data	44
Robustness test of the Machine Learning Model -- Multivariate Principal Component Analysis (PCA)	51
A Random Forest Regressor (RFR) was Used	53
Conclusions and Future Work	57

Essay 3: The *IntelliTurk* Tool for Evaluating Speech Intelligibility in Children: Software Tool

Introduction and Experimental Validation	61
Introduction.....	61
Motivation and Study Problem	61
Purpose and Research Approach	61
Limitations of Existing Approaches	62
Research Approach	63
Research Questions	64
Research Hypothesis	65
Limitations	65
Key Terms	65
Literature Review - Established Intelligibility Assessment Methodology	65
Systems Design.....	69
<i>Intellitürk</i> Solution Architect Design.....	71
Data Preparation and Analysis.....	74
Validation Study Research Method	76
Speech Samples	76
Children Speakers (Research Population)	77
Preparation of Speech Samples: Materials and Procedure	79
Experimental Design.....	79
Results.....	81

Hypothesis Tests	81
Econometric Regression Analysis	84
<i>IntelliTurk</i> Machine Learning Model.....	93
Mel-Frequency Cepstrum Coefficients (MFCC).....	93
Dataset	95
Data Preprocessing and Feature Extraction	96
The Deep Learning Model (DNN)	96
Lab Validation	97
Preparing and Testing the DNN	103
The Efficacy of the <i>IntelliTurk</i> Model Supported through DNN	104
Conclusions and Future Work	109
References.....	113
Appendix 2A. SCA Cupping Form.....	120
Appendix 2B: SCA Protocol, Roasting and Sample Preparation	121
Appendix 2C: SCA Protocol, Roasting and Sample Evaluation	122
Appendix 2D: SCA Protocol, Sample Evaluation Steps	123
Appendix 2E: Coffees’ Origin, Roasting Data and Their Final Quality Grading Scores	124
Appendix 2F: Coffee Green Parameters	125

Appendix 2G: Coffee Roast Parameters	125
Appendix 2H: Coffee Roast Profile (Excerpt) Time / Temperature.....	126
Appendix 2I: Coffee Roast Profile Time / Temperature Roasting Events	127
Appendix 2J: Drum Coffee Roaster Built at The Institute of Technology of Veracruz, Mexico	127
Appendix 2K: Integration of the Fuzzy Controller as Drum Temperature Regulator ...	128
Appendix 2L: Letter of Invitation for a Collaborative Effort Between the Coffee Technology Laboratory of the Food Research and Development of the Institute of Technology of Veracruz, Mexico	129
Appendix 2M: Stata Principal Component Eigenvectors	130
Appendix 2N: PCA Eigenvalues Scree Plot	130
Appendix 2O: Stata PCA Program (Script)	131
Appendix 2P: Eigenvector Values not Included in Comp1	132
Appendix 2Q: Eigenvector Values not Included in Comp2	132
Appendix 3A: Updated <i>IntelliTurk</i> IRB	133
Appendix 3B: Training of the DNN with the Spectrograms' Images	134
Appendix 3C: Validation of the DNN Upon Training with the Spectrograms' Images	134
Appendix 3D: Testing the DNN Deployed Model with a Classification of a Spectrogram	135

List of Tables

Table 2-1: Coffee Bean Attributes as Described by the SCA. Seven Attributes are Classified as Fuzzy Attributes (<i>Italics</i>), Four Attributes are Crisp Attributes (Normal)	38
Table 2-2: SCA Coffee Attributes Quality Scale	39
Table 2-3: SCA Coffee Final Grading Quality Scale	39
Table 2-4: Assumptions in a Fuzzy Control System Design	41
Table 2-5: Steps in Designing a Fuzzy Control System	42
Table 2-6: Seven Predictors and the Output Parameter (Score)	43
Table 2-7: Several Parameters Combined from Green and Roast Data	43
Table 2-8: Sample Roast Data	44
Table 2-9: Sample Data, Combining Origin and Roast	44
Table 2-10: Parameters Correlation Table	51
Table 2-11: Principal Component/Eigenvalues	52
Table 2-12: Experiments for Testing the Impact of PCA Components in the Machine Learning Model	54
Table 2-13: Eigenvector Omitted by Components	55
Table 3-1: Example List of Tokens and Confirmation Numbers Managed by <i>Intellitürk</i>	72
Table 3-2: <i>IntelliTurk</i> Slider Linguistic Terms Numerical Ranges and their Overlap	74

Table 3-3: Study Targeting Hypotheses Covering Three Treatments and Two Subjects Types	75
Table 3-4: Breakdown of the List of Fields (Data-Points) Captured During Experiments	75
Table 3-5: List of Speech Sound Subtypes	76
Table 3-6: Experiments Targeting Listeners Through the Amazon Mechanical Turk (AMT) Using Auburn University Experimental Design Platform (<i>IntelliTurk</i>)...	79
Table 3-7: DME Estimate Descriptive Statistics Inexperienced Listeners (group 1).....	82
Table 3-8: DME Estimate Descriptive Statistics Experienced Listeners (group 2).....	82
Table 3-9: Variance Ratio (VR) Test of Equality of Variance Inexperienced (Group 1).	84
Table 3-10: Variance Ratio (VR) Test of Equality of Variance Experienced (Group 2).	84
Table 3-11: Summary of the Three Predictors Used in the Models for Inexperienced Listeners (Group 1).....	85
Table 3-12: Summary of the Three Predictors Used in the Models for Experienced Listeners (Group 2).....	85
Table 3-13: Regression Analysis Results Inexperienced Listeners (Group 1)	88
Table 3-14: Regression Analysis Results Experienced Listeners (Group 2).....	89
Table 3-15: Spectrograms Images Resulting from Children Recordings	97

Table 3-16: Steps to Get a Deep Learning Model for the Prediction of DME Based on
Google Brain’s TensorFlow..... 98

Table 3-17: Portion of Resources Representing Child Recordings Used to Score DME 99

Table 3-18: DME Scores Given by Trained Clinicians and Collected by the
IntelliTurk.ML Windows-Based Application 99

List of Figures

Figure 1-1: Classes of Systems and their Associated Problems	78
Figure 1-2: Stablished Kinds of Reasoning and their Perspective toward Information ...	87
Figure 2-1: Coffee Roast Curve Plot, Y axis = temperature in Fahrenheit, X axis = time in Minutes	34
Figure 2-2: Roasting Profiler Solution	35
Figure 2-3: The Roasting and Cupping of Coffee Beans are two Separate Manual Processes	36
Figure 2-4: SCA Cupping Form's Strip, few Attributes of a Sample (Coffee Bean).....	37
Figure 2-5: Coffee Bean Quality Grading Using Linguistic Terms for Attribute Scoring Through User Interface	39
Figure 2-6: Fuzzy Expert System Architecture	40
Figure 2-7: A simple Fuzzy Logic Control System Diagram	41
Figure 2-8: Dependent Variable Score Distributed by Country	45
Figure 2-9: ANOVA Shows Statistical Significance of the Regression Model (All Countries).	45
Figure 2-10: Model Could Explain 14 Percent of the Variability of the data (All Countries).	45
Figure 2-11: Only the Intercept Shows Statistical Significance (All Countries).....	45
Figure 2-12: Fit Supporting Model Assumptions (All Countries).....	46

Figure 2-13: Scatter Plots (Colombia)	46
Figure 2-14: Model is not Significant (Colombia)	46
Figure 2-15: Individual Predictors Statistical Significance (Colombia)	47
Figure 2-16: Model Only Explain 21 Percent of Variability (Colombia)	47
Figure 2-17: Fit Supporting Model Assumptions (Colombia)	47
Figure 2-18: Model is not significant after removing outliers (Colombia)	48
Figure 2-19: Scatter Plots (Honduras)	49
Figure 2-20: Model is not Significant (Honduras).....	49
Figure 2-21: Model only explain 31 percent of variability (Honduras).....	50
Figure 2-22: Fit Supporting Model Assumptions (Honduras).....	50
Figure 2-23: Model is not Significant After Removing Outliers (Honduras).....	50
Figure 2-24: Model Only Explain 28 Percent of Variability (Honduras).....	50
Figure 2-25: Individual Predictors Statistical Significance (Honduras).....	50
Figure 2-26: Root-mean-squared Metric Used in the RFR Model	53
Figure 2-27: List of Parameters Available for the RFR Model	53
Figure 2-28: Excerpt of the Data fed Into the RFR Model	54
Figure 2-29: Setting up the Random Forest Regressor Model	54
Figure 2-30: PCA Component Loading (Comp1 and, Comp2)	56
Figure 3-1: <i>Intelliturk</i> Detailed Solution Architecture Diagram.....	70
Figure 3-2: User Intelligibility Assessment Response Input	70

Figure 3-3: <i>IntelliTurk</i> Experiment Design Administration Page.....	71
Figure 3-4: The <i>IntelliTurk</i> Supports Researchers to Engage Subject at their Lab by Using Confirmation Numbers.....	74
Figure 3-5: DME for Both Groups and All Treatments.....	83
Figure 3-6: Margins Plots (with 95% CIs) of Subtypes with Statistically-Robust Coefficients from Regression Model #1, Group 1.....	90
Figure 3-7: Margins Plots (with 95%CIs) of Subtypes with Statistically-Robust Coefficients from Regression Model #3. Group 1.....	91
Figure 3-8: Margins Plots (with 95%CIs) of Subtypes with Statistically-Robust Coefficients from Regression Model #1, Group 2.....	91
Figure 3-9: Margins Plots (with 95%CIs) of Subtypes with Statistically-Robust Coefficients from Regression Model #3. Group 2.....	92
Figure 3-10: Mel-Frequency Cepstrum Coefficients Logical Flow, it Converts Recordings into Spectrograms (Images).....	93
Figure 3-11: Math Works depicts Deep Learning as a Subset of Machine Learning, a Subset of Artificial Intelligence	94
Figure 3-12: Google Brain, Schematic TensorFlow Dataflow Graph for a Training Pipeline	98
Figure 3-13: Windows-Based Application Used by Trained Clinicians for Listening Child Recordings and Score their Intelligibly	100

Figure 3-14: Core Structure of a Deep Neural Network (DNN), Originally Proposed by Hinton in 2006	102
Figure 3-15: Testing the DNN Deployed Model with a Classification of a Spectrogram Image Employing a Windows-Based Application	104
Figure 3-16: Model Including Very-Difficult and Very-Easy (Group 1)	106
Figure 3-17: Model Including Difficult and Easy (Group 2)	106
Figure 3-18: Model Including Very-Difficult, Difficult and Medium (Group 3)	107
Figure 3-19: Including Medium, Easy and Very-Easy (Group 4)	107
Figure 3-20: Model Including Very-Difficult, Difficult, Easy and Very-Easy (Group 5)	108
Figure 3-21: Model Including Very-Difficult, Difficult, Medium, Easy and Very-Easy (Group 6)	108

List of Abbreviations

AMT	Amazon Mechanical Turk
API	Application Programming Interface
ART-MAP	Adaptive Resonance Theory Method
CQI	Coffee Quality Institute
CSS	Cascade Stilling Sheets
DNN	Deep Neural Network
DME	Direct Magnitude Estimation
IHCAFE	Honduras Coffee Institute
ML	Machine Learning
MLP	Multilayer Perceptron
EV	Expression Evaluator
FSI	Fuzzy Inference System
GNU	General Public License
I-VAM	Infrastructure Vulnerability Assessment Model
NN	Neural Network
MEDAS	Direct Magnitude Estimation Data Analysis
MFSIS	Mamdani Style Fuzzy Inference System
MVC	Model View Controller Design Pattern
PCA	Principal Component Analysis
POC	Proof of Concept
RFR	Random Forest Regressor
UI	User Interface

SCA	Specialty Coffee Association
SCAA	Specialty Coffee Association of America
SaaS	Software as a Service
SGDR	Stochastic Gradient Descent with Restarts
SMEs	Subject Matter Experts
SVM	Support Vector Machine
2FS	Type-2 fuzzy set

Essay 1: Designing Machine Learning Models Leveraging Linguistic Scoring Uncertainty (De-erroring)

Introduction

This chapter will provide an approach for quantifying judgment, the foundation of human knowledge representation and reasoning when making decisions in the presence of vagueness, imprecision, and uncertainty. Two highly complex systems whose behavior is not well understood will be given an approximated solution (e.g., speech intelligibility, and coffee roasting and judging), and we will model their uncertainty using numeric and linguistic scoring techniques, fuzzy logic and machine learning.

The preponderance of cognitive and heuristic biases in human decision making is well established, Montebellier & Von Winterfelt [3]. Moreover, in light of recent developments in machine learning designed to mimic human behaviors, the next frontier for both machine learning and the decision sciences will address the degree to which human decision biases (including cognitive errors and heuristic biases) can be mitigated using machine learning methodologies. In short, if machine learning can be utilized to replicate human decisions along with all of their biases and flaws, can machine learning also be used to reduce or even correct those biases? Many research domains and conferences within the fields of decision sciences, psychology, and human factors engineering identify methodologies for addressing cognitive and heuristic biases with applications as wide reaching as terrorism interdiction, jury selection, and medical diagnoses. This chapter will peer into tomorrow's frontier by addressing the imprecision embedded in human judgment and decision making relative to the next generation of machine learning approaches [4] [5] [6] [7]. The aim is to integrate the naturalistic mechanisms of human decision making with machine learning capabilities.

Beyond the Organized Simplicity Perspective

Halpern, 2017, states, “whichever approach is used to model uncertainty, it is important to be sensitive to the implications of using that approach” [8, p. 55], and Ezell considered this when designing I-VAM, a model that was successfully employed to score clean water plans [9]. In this context, Halpern clarifies that the uniqueness of an application requires an approach that is appropriate for handling its uncertainty (call for choosing a representation). Halpern’s observation provides a solid leaning perspective for this research, in particular when no quantitative information is available and a tool, system, or application does exist to offer the necessary accuracy. Hence, as a solution is designed and built, the appropriate approach to handle its uncertainty must be handled with sensitivity [8] [10] [11]. An analogous case that we see in nature is how humans cope with decision-making mechanisms. This is applicable to banking, investments, risk management, and food consumption decisions [3].

The current information systems marketplace is defined by many developed and maintained arrays of information systems that produce volumes of collected data. This observational data has become the raw material for data scientists and others interested in developing models to make inferences.

In human cognition, it is well established that as randomness and complexity increase, so does the organized complexity within the randomness of the system, and vice versa. Generally, in software engineering, individuals have the tendency to operate within organized simplicity, a comfortable place where our system design is based on the known states of inputs and pre-determined outputs (i.e., deterministic systems in bank account management applications). We as researchers want to reach into disorganized complexity and extract tractable solutions when the

output of the system is not well understood from the initial states of its inputs (i.e., non-deterministic systems).

According to Klir, 2017, a fully operational theory, such as the uncertainty-based information theory or any other theory targeting the uncertainty of some conceived type, addresses uncertainty issues using four levels [12]. Level one provides the mathematical formalization. Level two consists of the calculus needed to properly handle this uncertainty. Level three features the ways to measure the amount of relevant uncertainty in any formalized situation covered in the theory. Lastly, level four addresses the need to develop methodological aspects of the theory, including procedures for making various uncertainty principles operational within the theory [12].

The broad array of available methodologies, such as Waterfall, Agile, Rapid Application, DevOps, etc., each have their strengths and weaknesses [13], and they are mainly driven by a set of practices geared toward adopting an organization size and field of specialization. All of these methodologies facilitate the task of assisting software engineering in growing its organized simplicity. Figure 1-1 shows three types of systems and the associated problems that require those four levels of the theory. Software development features applications within the scope of organized simplicity.

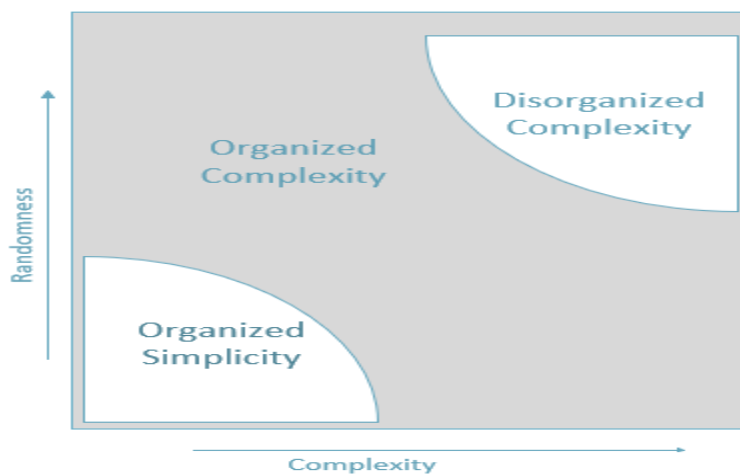


Figure 1-1. Classes of systems and their associated problems [12].

Unlike a conventional system analysis that first poses a model, the uncertainty contained in both the inputs and outputs of a system are employed when formulating the system's structures. In this chapter, our research focuses on the formalization of human imprecision, specifically the uncertainty carried by Scoring Numeric Processes driven by probabilistic methods. Additionally, we discuss the formalization of human imprecision and the uncertainty introduced (a) by tools used by humans, and (b) by the translation of human perception of any stimuli into a derived measurement despite the complexity of the phenomenon. Furthermore, we explain that transitioning from numerical to linguistic scoring methods, or by combining both numeric and linguistic methods, as shown by Livio et al., 2020, allows for the measurement of imprecision.

The debate concerning how uncertainty should be represented when modeling natural phenomena is an ongoing and heated one [8, p. 12]. Some researchers believe that there is only one suitable mechanism, the probability theory, to study numeric uncertainty [12, p. 11]. However, although the probability theory (a mature and well-understood theory) has come a long way in providing an appropriate modeling mechanism that addresses uncertainty, in particular numeric uncertainty, probability theory faces some challenges (e.g., the fact that a model's numbers are not always at hand). Because probability theory is based on numbers, it compares two events in terms of probability by considering scenarios when both events have the same probability of occurring or when one event is more probable than the other.

Moreover, another set of arguments support, under certain assumptions, that the only rational way to represent uncertainty is Dubois' possibility theory because "the notion of probability seems less flexible than of possibility" [14, pp. 10-12].

Additionally, classical logic has a serious limitation in that it cannot cope with the issues of vagueness and uncertainty, which exists in most modes of human reasoning [15] [16, pp. 265-

277] [17, pp. 8-10]. Contrary to Rutherford's dictum: "qualitative is nothing but poor quantitative," computer science and its applied arm of software engineering are supporting artificial intelligence in considering quantitative attributes with nothing but limited knowledge of the qualitative [18].

Here we focus on those non-deterministic systems where the uncertainty is evident and pertains to the purpose for which the system was constructed [12, pp. 4-7]. Software engineering involves designing and developing solutions, driven by system requirements and derived business logic, and employing a subset of methodologies and tools. These methodologies and tools carry uncertainties that find their way into the resulting data. Software engineering and software architecture see systems as a relation among states of given entities. Hence, each system's relation targets some unknown states of given entities with the purpose of providing values from the basis of the known states of other involved entities. However, it is not always possible to deterministically set those unknown states from those that are known [10] [12] [19].

Furthermore, extensive research supports that "heuristics are quite useful, but sometimes they lead to severe and systematic errors" [20, p. 104]. Shah and Oppenheimer have developed an effort-reduction framework [21]. They studied heuristics (i.e., simple processes that replace complex algorithms), as it serves as a diminisher of the amount of energy a decision-making task requires [21]. However, from the perspective of software engineering and computer science, in particular when eliciting human knowledge (i.e., designing membership functions), researchers must move beyond established facts and consider how those complex algorithms represent complex systems (i.e., the most exact mathematical representations of a phenomena are context-dependent [19, pp. 34-25]), modeling the underling phenomena associated with the human decision processes that are engaged either by judgment or choice). Furthermore, Ezell studied the

importance of assessing a system's conditions (i.e., vulnerability, uncertainty) within the "context of a scenario" [9], and introduced the model of infrastructure vulnerability assessment (I-VAM). I-VAM is a great example of knowledge engineering, as it requires subject matter experts (SMEs) to derive value functions and weights along with the assertion of protection measures of the proposed system. For example, I-VAM was used to score a clean water system that served a residential community and supplied approximately 2 million gallons of water per day [9, p. 575]. Ezell reported that "significant changes to the weights (provided by SMEs) may change the score, but just as importantly, inform where improvements could be made to reduce vulnerability in the system" [9, p. 581].

This research aims to contribute toward level four proposed by Klir. In this work, we are not embracing organized simplicity nor "simple processes that replace complex algorithms" [12, pp. 4-5]. Instead, this study takes a look at disorganized complexity from the perspective that organized complexity simultaneously feeds on a blend of discussed theories such as Possibility Theory and Fuzzy Set Theory (i.e., Fuzzy Logic) and chaotic complexity. However, this work does not intend to increase simplicity, but, as disorganized complexity is reached, will instead build a pathway to formalize the uncertainty involved in each of the steps (the ones clearly handled through software engineering processes). Here, uncertainty and probability are not equal, and uncertainty has been liberated from its probabilistic confines [12].

So far, we have defined uncertainty, vagueness, elasticity, and complexity from software engineering perspective. Next, we describe two terms that support the conceptual contributions of this chapter.

Bias and Error

Technically, error has a direct impact on the variance. However, error does not affect the mean [22, p. 19]. By recognizing the fact that all measurements carry error, we have a solid ground on which to develop strategies that compensate it [23, p. 189]. Additionally, an error in a consistent direction becomes bias, a systematic error with “an inherent tendency of a measurement process to favor a particular outcome” [23, p. 190]. This is unfavorable because it leads to conclusions other than the actual true point estimates in the object of our measurement, as is the case when bias affects the mean [22, p. 10]. Koehler and Harvey suggest a set of tools to address the bias when domain experts (i.e., practicing physicians) engage in decision making through their inherent judgment, what psychologists call “the optimistic overconfidence perspective” [20, p. 209-201]. Their study considers this a natural perspective that over-predicts when outcomes are desirable and under-predicts when outcomes are un-desirable. However, their study observed other natural perspectives, such as confirmatory bias and positive and negative moods bias, where those outcomes with higher probability resulted in over-prediction and those with less likelihood of occurrence were under-predicted by experts [20, p. 209].

De-biasing

Koehler and Harvey suggested tools to correct bias (e.g., sunk cost bias, recency bias, etc.) in what they called “a debiasing problem.” Their effort involved the study of a set of tools targeting these biases to improve decision-making [20, p. 412]. For example, utility assessment is one of the most influential tools in medical decision-making research. It supports researchers, policymakers, and individual clinicians. Clinicians use it to consider pre- and post-treatment health states and the side effects of surgery in an individual patient or a group of similar patients. Additionally, when

debiased, an ethical decision of whether or not to approve the payment for an intervention can be made, especially when placed on the shoulders of a policy maker who is well aware of the normative regulations and specifications that drive the system's requirements, as dictated by stakeholders. Hence, increasing the patient's quality of life would be the predominant factor driving the decision [20, p. 608-610].

De-biasing patients' health decisions gives one the ability to easily and accurately improve their decision-making process and has many applications, such as. Thus, the need for designing decision aid tools (i.e., for customized treatment recommendations) based on the data pertaining to an individual patient's preferences for health states.

Two recent applications incorporate this concept: a hearing aid solution that incorporates both hardware and software components (Starkey Livio) and features a telehealth service that the care provider to "troubleshoot and improve the patient experience by delivering programming adjustments directly to a patient's smartphone and hearing aids with no need for them to stop in" [24]; a customizable, demographically profile-based diet (Noom) that claims "Quick fix diets are a thing of the past, and behavior change is the way to the future" [25].

De-erroring

In this exposition, bias is an error in a consistent direction and can be corrected with some tools. Nonetheless, researchers have also suggested some tools to correct error as it affects the decision-making process (e.g., availability bias, correlation error, gain loose error, over confidence) [3] [20].

Software engineering is a systematic discipline concerned with all aspects of software production, from the early stages of system specification to maintaining the system after it has

gone into use [13, p. 10]. It is centered around organized simplicity, and as it approaches complexity, randomness (error) is introduced. Figure 1-1 shows an ample region in its center titled “Organized Complexity.” Regarding this region (which science has not explored enough), Klir remarks that problems often involve “a considerable number of variables” [12, p. 4]. Moreover, Klir indicates that at this organized complexity region is where high computing power is needed. However, it is not enough to make substantial progress [10] [12, p. 4]. It is there, at the organized complexity region where radically new approaches are needed to formalize this “broad concept of uncertainty” [12, p. 4-5].

We have identified that when human decision-makers are forced to use a numeric system for evaluation in contexts where numeric values are foreign or not naturalistic, this leads to error (not necessarily bias). Therefore, considering that psychologists are proposing the use of debiasing tools and that a vast number of tools are available within software engineering and computer science, this research anchors its aims in a de-erroring tool, a tool that can correct or mitigate error. Moreover, this research considers business applications that target daily decision-making processes, where decision-makers often choose precision with unknown bias (that is not benign) instead of highly imprecise measurements (that are manageable) with random error [23].

Recent Applications of Machine Learning Supporting Human Decision-Making

Data are the fuel of machine learning techniques. Learning is possible because of data [26], and as data comes through the tools that integrate decision-making processes, it incorporates human interaction.

Researchers and entrepreneurs from various disciplines are reaching out to computer scientists, data scientists, and software engineers to help them build less error-prone applications. These applications are used by humans daily for basic and complex decision-making processes, such as driving a car along a busy highway, grading school assignments, and asking a patient for their level of pain (i.e., by presenting a scale from 1 through 10). Human travel locally, nationally, and internationally demands the need for real-time information on catastrophes and pandemics, such as accidents, hurricanes, and COVID-19. In particular, as we follow the COVID-19 pandemic news locally and globally, it is clear that the human decision process is engaged, however, the decision results do not match the expectations of involved authorities and scientists. This fosters another question: is the decision-making process at the level required to handle this event?

Furthermore, most non-deterministic systems collect an increasing amount of data (e.g., aerospace records, medical records, daily trading, money exchanges, investments, law suits, retirement information, social security transactions, traffic patterns, etc.). For example, currently the field of space exploration features anomaly detection in spacecraft telemetry and exoplanet detection. These unique applications allow engineers and scientists using terabytes of data collected from space missions to improve their decision-making processes [27] with machine learning techniques, specifically convolutional neural networks (or CNNs). This is done by carefully preprocessing the data, setting the network parameters, and addressing underlying physics constraints while constructing the system [27]. The output of one system eventually becomes the input for another application, and, in most cases, data changes through a flow of

decision-making steps. Moreover, this data, when properly aggregated and used with machine learning techniques, will predict new events, classify new diagnoses, and look for cluster patterns not visible to the human eye. All of this will contribute to the ultimate goal of assisting humans in our decision-making process, a core element of our human nature.

Supporting Theories in the Literature

Possibility Theory

A theory has emerged and matured to the point of providing enough foundation to deal with these systems in the most qualitative way possible. Known as Possibility Theory [28] [29, pp. 531-571], it is a modern and simple mathematical theory of uncertainty that is sufficiently used when dealing with incomplete information or imprecise probabilities. Zadeh developed the idea of Possibility Theory from his previous popular work on fuzzy sets and fuzzy logic, which were first postulated in 1965. See [30] for Zadeh's remark: "imprecision that is intrinsic in natural languages is, in the main, possibilistic rather than probabilistic in nature."

Possibility Theory supports reasoning under uncertainty, and has taken the role of addressing some issues that classical probability-based methods cannot solve. For example, the lexical elasticity of the predicate small and large, perfectly handled with fuzzy set theory, dealing with the possible rather than probable values of a variable with the possibility being a matter of degree (elasticity) [28].

Additionally, research indicates that the unique relationship between fuzzy set theory (or fuzzy logic) and the possibility theory is the result of inconsistency between available knowledge and what is considered possible (a fuzzy set type) from current observations of the phenomenon being studied [31, pp. 12-13]. This has proven useful in the scope of statistical reasoning, like

when providing a solution using information containing uncertainty due to variability or vagueness and incomplete or imprecise information.

Why is Possibility Theory a part of this research and how it is connected? As part of our research, we have employed fuzzy logic. As the literature indicates, fuzzy logic is a method to formalize the human capacity of imprecise reasoning (approximate reasoning) [32] [33]. In addition, a clear distinction between fuzzy set theory (fuzzy logic) and the probability theory is well presented by the concept that, the kind of uncertainty associated with the impression present in the prescription of the boundaries of a set, it could be well represented by a possibility distribution [14] [28].

In a recent work, Dubois et al. published in 2016 [34, pp. 24-6], remarks that possibility theory is one of the main theories for reasoning under uncertainty due to incomplete information. This theory is described as “a flexible framework for merging information because set based fusion modes can be directly extended to fuzzy sets representing those entities of incomplete information in a gradual way.” While possibility theory represents imprecision (in the form of fuzzy sets), it also quantifies uncertainty through the measure of possibility and necessity [14] [28]. Moreover, it illustrates conceptually that a possibility distribution is, in fact, a fuzzy restriction performing as “an elastic constraint on the values that may be assigned to a variable” [31]. In fuzzy logic, a membership function representing a linguistic term, which maps a possibility distribution also called possibility measures by Zadeh [14] [15] and [30]. This makes the connection between these two frameworks, the Fuzzy set Theory and the Possibility Theory.

Uncertainty-based Information Theory

The uncertainty carried by a non-deterministic system rarely is a result of randomness, the producer of meaningful statistical averages [12]. Uncertainty-based Information Theory challenges the early assumption that this uncertainty only could be dealt with the resulting framework of the classical probability theory, a perspective that “uncertainty and probability are equal”. A thought that has been challenged by two important mathematical generalizations: a generalization of the classical measure theory and a generalization of the classical set theory, both enlarged substantially the framework for formalizing uncertainty by making possible to conceive new uncertainty theories departing from the classical probability theory [12, p. 6]. The Uncertainty-based information, uncertainty involved in any problem-solving situation is a result of some information deficiency pertaining to the system with which the situation is conceptualized” Klir J. George, 2006 [12] and [17]. This a theory that distinguishes the information conceived in terms of uncertainty reduction from other conceptions of information. This theory perfectly supports the modeling of non-deterministic systems, systems not based on the conception of information in human communication and cognition, or the algorithm conception of information. It moves away from systems where the amount of information needed to describe an object is measured by the shortest possible description of the object in some standard language, a language that resembles the common-sense of what is perceived during the interpretation of the system’s involved information.

However, formal treatment of uncertainty-based information derives from two classical roots, one comes from possibility [28] and the other from the notion of probability [12], [17] and [31]. Why is Uncertainty-based Information Theory a part of this research and how it is connected? A core element of our research is the formalization of human impression. We do depart from the perspective that indeed, information reduces uncertainty, which gives me the foundations Instead,

it inspires me to contribute toward the understanding of human communication and cognition explained from the perspective of both deterministic or not deterministic systems.

In summary, Uncertainty-based Information Theory strives to develop the capability to deal formally with any type of uncertainty and the associated uncertainty-based information that one can recognize on intuitive grounds [12]. Hence, connecting uncertainty-based information theory with fuzzy set theory which at the same time is connected with possibility theory as remarked above.

Fuzzy Logic

For this section, we start with the question: Why is fuzzy logic a part of this research and how it is connected with it?

Fuzzy logic, the first theory with a theoretical application of the uncertainty and vagueness, supports the generalization of classical measures by leaving behind sharp boundaries between sets as seen by contemporary science [35, p. 481]. Furthermore, fuzzy logic has proven to be a superset encompassing classical bivalent logic, generally presumed to be the principal pillar of science. Fuzzy logic's propositions are not required to be either true or false, but may be true or false to different degrees. This aggressive approach (the degree of being either partially true or partially false, elasticity) has shown that some bivalent logic laws do not hold any longer, e.g. the law of excluded middle and the law of contradiction [12, pp. 418-420].

Research has shown that fuzzy logic offers a novel approach when other mathematical models (i.e., the linear regression statistical model) are not available for processes where human experts do well at mastering. Moreover, in order to provide a foundation for human knowledge representation and reasoning in the presence of vagueness, imprecision, and uncertainty [29].

Fuzzy logic, in particular, the fuzziness of a fuzzy system is based on the fact that such a system handles very well the imprecision in the underlying model parameters, it deals with it rigorously and predictably [36, pp. 12-13].

Additionally, fuzzy logic does not see imprecision as the result of missing data, cloudiness in the knowledge base, the probability that one event occurs, or sloppiness in the model design, for fuzzy logic, “imprecision rest in the natural and real-world imprecision associated with nearly all-natural phenomenon” [36, p. 12].

More, fuzzy logic consists of a set of mathematical principles e.g. level one and two as suggested by Klir [12] and proved by Zadeh [15] [30] [37], for knowledge representation based on degrees of membership rather than what classical binary logic features off a crisp membership. With the power of fuzzy logic, now classical binary logic is considered as a special case of multi-value fuzzy logic [38, p. 89].

Hence, we can see fuzzy logic as a method of encoding and using imprecise information. However, its usefulness results when it works combined with analytical methodologies as suggested by Klir as well [12] [17], machine reasoning techniques, and the decision support apparatus inherent of conventional expert systems [17], [36] and [39]. This because when fuzzy logic is used, the underlying fuzzy sets define the semantics of the model as well as the precise relationship between data points (models’ state) and the set membership (elasticity).

Because fuzzy sets are always concrete and deterministic, the only actual imprecision or fuzziness is associated with the very high level of the structure represented by the fuzzy sets themselves [36]. Ross puts it this way “Natural language despite its vagueness and ambiguity, is the vehicle for human communication, and it seems appropriate that a mathematical theory that

deals with fuzziness and ambiguity are also the same tool used to express and interpret the linguistic character of our language” [29, p. 118].

Deep Learning

Deep learning is inspired by the architectural depth of the brain, researchers wanted for decades to train deep multi-layer neural networks. No successful attempts were reported before 2006. Researchers reported positive experimental results with typically two or three levels (i.e. one or two hidden layers), but training deeper networks consistently yielded poor results. In 1993, Vapnik and his co-workers developed the Support Vector Machine (SVM, a shallow architecture). Some exceptions continued, for example, convolutional neural networks were reported by Yann LeCun in 1998. However, these approaches led to a digression, in the 1990s, many researchers abandoned neural networks with multiple adaptive hidden layers because SVMs worked better, and there were no successful attempts to train deep networks. Until 2006, deep multi-layer neural networks had not been successfully trained. Since then, several algorithms have been shown to successfully train them, and now experimental results have shown the superiority of deeper vs less deep neural network architectures. Deep learning methods focus on learning hierarchies with features from higher levels of the hierarchy formed by the composition of lower levels [5]. Also, deep learning employs a wide scope of learning methods ranging from neural networks with many hidden layers [4] to graphical models handling several hierarchies of hidden variables.

Deep learning architecture (architectures that are composed of multiple levels of non-linear operations, such as neural nets with many hidden layers) brought some advantages and challenges. Moreover, theoretically speaking, the deep learning architecture allows us to model some complicated functions not efficiently represented (in terms of the number of tunable elements) by

architectures that are too shallow and, it might be able to represent some functions otherwise not efficiently representable.

Such functions that can be compactly represented by a depth k architecture (multiple levels of non-linear operations) might require an exponential number of computational elements to be represented by a depth $k - 1$ architecture. Hence, computationally there is no need for many elements in the layers, and statistically, poor generalization may be expected when using an insufficiently deep architecture for representing some functions [7].

Why is Deep Learning a part of this research and how it is connected with it? As part of our research, we have to acknowledge some limitations of fuzzy expert systems, whereas they are very useful in situations involving highly complex systems whose behaviors are not well understood and in cases when an approximate, but a fast solution is derived [29] [36]. Nonetheless, some scholars have remarked that because fuzzy systems are primarily used in deductive reasoning, where the specific is inferred from the general (deductive reasoning, also called shallow reasoning [29, p. 8]), but not do well when modeling systems based on inductive reasoning (inferring the general from the particular) or deep reasoning, requiring models to capture those processes, core players of the mother nature that produce the phenomenon we observed such as listening to a child in order to determine its intelligibility or seeing a human driving a car and being able to model these phenomena.

Theoretical Problems to Address

Two major problems have been addressed, an industrial problem (coffee) and a social one (speech pathology). As studied by Jorge Klir [12] and proposed by Jerry Mendel [17] [20] “When dealing with real-world problems, we can rarely avoid uncertainty” [19]. Nonetheless,

there is a core distinction when considering measurements and perceptions: “measurements are crisp numbers where perceptions are fuzzy numbers (represented linguistically)” [19]. This is based on the original work of Zadeh, on a paper presented in 1975 where he provided the foundation of a qualitative approach successfully applied on modeling complex systems behavior by using linguistic instead of numerical variables [15] [16] [32] [40] [41].

Industry Problem: Roasting and Judging (Determining Quality Scoring) Coffee Beans

An industry-standard protocol (roasting high-quality coffees beans and its impact into its final grading, based on their attributes, see Appendix 1B, 1C) [42] [43]. The approach. First, coffee beans from various countries were profiled. These coffees were roasted; batch profiles were recorded focusing on various parameters (e.g. charge temperature, turn time and temperature, first and second cracks’ time and temperature and total roasting time). Second, a Fuzzy Expert System was used to determine the final quality grading of these coffees as indicated by the SCA standard protocol (see Appendix 1A, 1B, and 1C). Third, these two datasets were combined by matching each of these coffee beans’ country of origin, quality grading score with their roast profiled batches (see Appendix 1E).

As the coffee beans were profiled, we captured domain knowledge and, we used the Experts’ knowledge captured through a Fuzzy expert system [33] while quality scores were given. In essence, a subset of soft-computing, Fuzzy Logic was employed to capture the underlying phenomenon such as the one inherent in the process of judging coffee beans by following the industry-standard protocol.

This chapter includes a solution to gather all the data needed for the tracking interrelationships, interdependencies, and co-relationships between the roasting process (getting

raw coffee bean ready for human consumption) and the cupping process (determining the coffee quality grade). Hypothetically, this research leads the way toward the design and development of a fuzzy controller to be integrated into a coffee roaster used to roast high-quality coffee beans while machine learning techniques are employed throughout the simulation process, profiling coffee roast curves and their key parameters (critical cut points along the roast curve, drivers of the cupping results targeting the final coffee bean quality grading).

Furthermore, other roasting parameters, ambient parameters (e.g., humidity) hypothetically impact the roasting process, with unknown and unmeasured impacts on the results (cupping to determine quality) of the coffee, they were also captured.

Social Problem: Ascertaining Children’s Speech Intelligibility

A problem currently in the speech pathologic field has been addressed. The need for having a standardized tool to ascertain children’s speech intelligibility. An exhaustive literature review was conducted and a well-established method known as Direct Magnitude Estimation (DME) was chosen. The result, a second application was built implementing the DME as a platform to design and conduct experiments (*IntelliTurk*).

Using the *IntelliTurk*, experiments have been designed to capture direct magnitude estimation, DME measurements from experienced or domain experts (speech pathologists, technicians), and inexperienced listeners using the *IntelliTurk*’s platform with three treatments, linguistic, numeric and both combined.

The scoring mechanism with less variability (error) is linguistic hence, it is proposed as the default method of evaluating and scoring speech intelligibility. Moreover, in order to employ machine learning techniques. A subset of recordings capturing the speech of sixty-two children

selected to articulate more than two hundred words was used. Another application was designed, coded, and presented to trained clinicians. These clinicians were instructed to listen to a portion of these recordings and then ascertain the DME using the new default scoring method.

When given audio samples in a computer-readable format (such as a .wav file) of a few seconds duration, holding a child reproduction of the recorded words used by the experts to linguistically measure the DME, we want to be able to determine (predict) the DME associated linguistic terms with those recordings. The goal, to have a prediction accuracy of at least eighty percent (80%). The following terms, a set of predictable, classifiable labels for the DME. The idea is to have a linguistic term for the final classification of the DME, initially suggested terms for the DME are: Very Difficult, Medium, Easy and Very Easy.

Deep neural networks (DNNs) have become very popular for image classification problems, they do it with high accuracy and at scale; our research employs DNNs for the prediction of the (DME), and to model the relationship between roasting coffee beans and its impact in their quality, we employed a Random Forest Regressor (RFR).

Theoretical Approach to Solve Proposed Problems

We presented above that heuristic and bias research have combined perceptual principles with the psychology of thinking and reasoning. This equipped us with a new perspective on judgment under uncertainty and, to consider, based on irrefutable evidence that humans' reasoning and decision-making capabilities, though certainly remarkable, are prone to systematic errors (i.e. bias) [20, p. 105]. These contributions, once considered from the software engineering and computer science perspective, offer us an opportunity to contribute. Hence, our de-erroring

approach if used, for example when developing applications supporting human decision-making, spanned across numerous domains, when utilizes machine learning, it can (hypothetically) disrupt the systematic propagation of the error, hence by de-erroring, it de-biases as well. Here we discuss two such domains based on existing research as an application. Therefore, we have conducted experiments to test and improve human decision making with human subjects and, by developing machine learning models trained and validated with data captured through non-deterministic applications (i.e., systems designed to model underlined phenomenon). Livio et. al. [32] [33]. Furthermore, when human decision-makers are forced to use a numeric system for an evaluation in contexts where numeric values are foreign or not naturalistic, this leads to error (not necessarily bias, see Appendix 2A, 2B, 2C, and 2D).

When a numerical value is attached to a logical formula as the result of logic-based knowledge representation such as requirement engineering, business intelligence, autonomous systems, software engineering including medical or engineering decision support, this value can be explained in many different ways according to its semantics [44]. Researchers agreed with a subset of common justifications for this value, e.g. belief degrees preference degrees and trust degrees [11] [44].

Ma introduced a numerical characteristic function for each knowledge-based approach and concluded with the existence of one to one correspondence between a set of combination rules of belief functions. It is observed in [45] that researchers are framing combination rules as the core of merging from “ancient concepts of non-additive probabilities to the modern concept of belief functions”, supporting with evidence that the degree of belief is the result of comparing evidence to knowledge about chances governing the truth founded in a philosophical perspective [45, pp. 38-41].

As a result, combination rules such as Dempster's rule [45], Smets's rule [46] and merging methods support the use of weighted knowledge base representations [47], by proposing a knowledge merging operator capable of solving the conflicts among the knowledge bases. However, these perspectives keep on calling for more exploration to be done. Mainly because researchers have noticed that sources may be partially dependent, but not completely, leaving room for wrong interpretations of the data models and the uncertainty they might carry [48].

On the other hand, uncertainty in information takes different shapes and it could be derived from ignorance, a variety of stochastic sources, from the inability of conducting adequate measurements, from lack of knowledge or from the fuzziness carried on our natural languages as we express certain phenomenon [29, pp. 13-14].

The various sources of uncertainty can either be study separately or combined. For example, as the imprecision or vagueness associated with data increases, overwhelming Fuzzy type-1 methods; Fuzzy type-2 offers a more robust approach to handle these uncertainties [12] [17] [19]. This is done by taking into account the inherent relationship encountered in precision, data, and technique [8] [49].

Nonetheless, other researchers had remarked that a distinction should be made, "in order to account for the underlying realities" [29, pp. 8-9] between employing mathematical models when studying the observed data and, those hidden but important elements that produce this data. The reason, models that cover only part of the phenomenological aspects, the part of the behavior being studied, are by definition shallow models. Models that do not perform well due to their lack of knowledge of the inherent processes that by nature account for these phenomena [8, p. 11].

Our research sees these uncertainties from an angle of problems we solve in software engineering. These problems are framed into imprecise and precise, formalizing their imprecision

as data is modeled from numbers into words and, from words into perceptions. However, the techniques we used have their part in the encountered uncertainties. Coupland and John remarked that “problems that contain precise data, should not be expected to exist.”

For example, mathematical modeling, type-1 fuzzy sets, and logic or type-2 fuzzy sets and logic, have been studied and proven to capture with different precision, the uncertainty carried by the perceptions we embrace as we interact with the underlined systems [19, pp. 246-248].

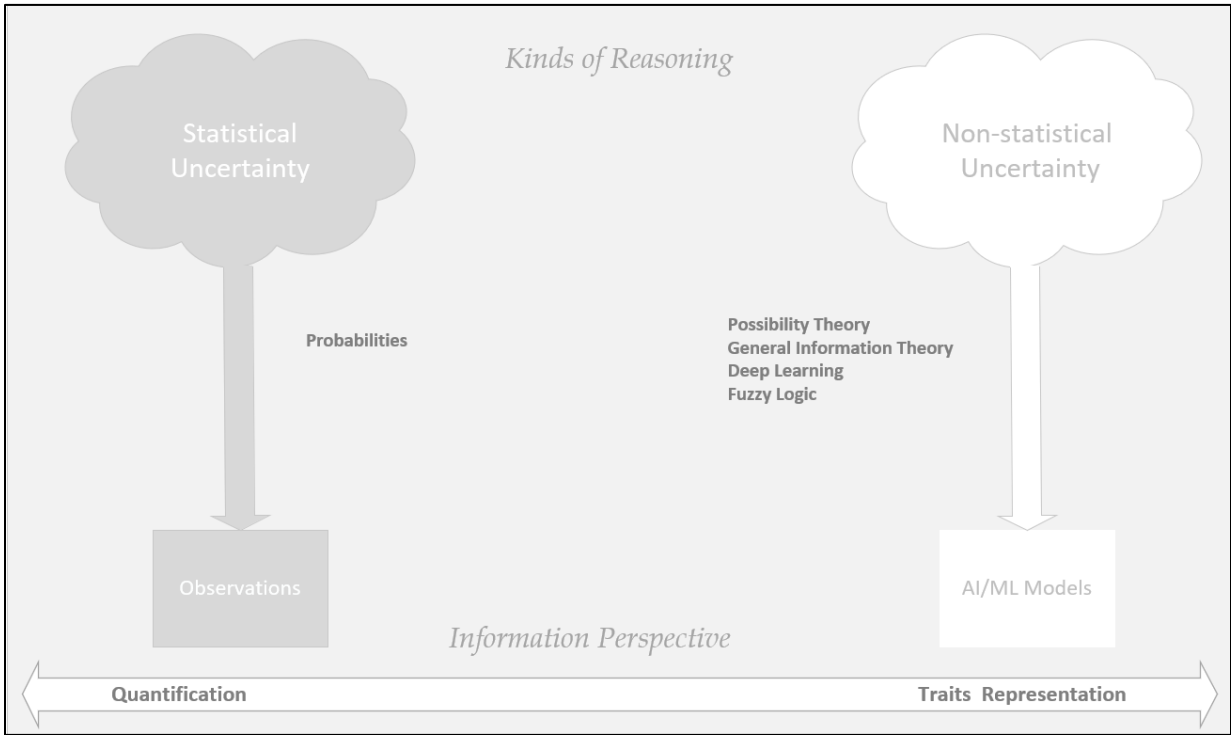


Figure 1-2. Kinds of Reasoning and Information Perspective

Figure 1-2 depicts the established kinds of reasonings and their information’s perspective. Non-statistical uncertainty is equivalent to fuzziness, it is an inherent property of a system and it cannot be altered or resolved by observations [35, p. 10]. Moreover, “the amount of fuzziness is not connected in any way to the quantification of information, it is an important trait of information representation” [12, p. 321]. On the other hand, Didier and Prade remarked that “probability theory

seems to be too normative a framework to take all aspects of uncertainty judgment into account” [14, p. 5] and even though the calculus of formal logic seems to be appropriate, nonetheless it falls short as it distinguishes just between true and false propositions whereas humans in a natural way, and efficiently cope with “highly subjective partial beliefs and drawing reasonable conclusions from distorted, hazy information of restricted reliability” [49].

Additionally, the use of numeric systems results in biases that can be improved upon with more human-friendly linguistic techniques that are now enabled by the use of new machine learning technologies [19] [35] and [38]. These biases (biases due to the use of numbers) are called item response bias, proxy bias, range insensitivity bias, and a full class of stimulus-response biases known as scaling bias (which includes contraction bias, logarithmic response bias, range equalizing bias, centering bias, and equal frequency bias).

In this perspective, [35, pp. 451-480] presents grounds to develop software products that model human reasoning, approximate reasoning, it has been framed from the combination of fuzzy logic with the fuzzy set theory where an element belongs to a set to a degree, indicating the certainty (or uncertainty) of membership, in a realm of characterizing domains by linguistic terms, rather than by numbers. Moreover, two-valued set theory enabled the development of exact reasoning systems when coupled with traditional Boolean knowledge [35].

In an effort to formalize each of these uncertainties, uncertainty-based information depends on the mathematical theory from which the conceptualization of its problem-solving situation is based. Hence, a search for a mathematical model to represent these situations is an undergoing endeavor whose influence has rippled the ultimate goal of generalized information theory: “to capture properties of uncertainty-based information formalized within any feasible mathematical

framework” [17, p. 246]. However, as remarked by Kruse et al., “the basic intention of any model is to reflect properties of the real world, i.e. to enable the prediction of a system’s behavior in the real world” [49, pp. 2-4].

Furthermore, cognitive scientists have studied those conceptual patterns and mental images from which humans conceptualize process, and store input knowledge domains. For example, Borghis [50] made clear that perception and action are not separate and sequential processes but that they are deeply interwoven. Other disciplines like engineering profoundly use numerical quantities. Where the latter operates on solid theories, the former acknowledge that despite the vagueness and ambiguity in natural language, “the shape of our thoughts” [37]. Human communicating in a common language have little trouble in basic understanding, supporting the idea that when modeling the human thought process as expressed in our communications with one another, this model must emulate our natural language [29, pp. 239-141].

Furthermore, employing machine learning techniques has proven to produce accurate models. Models that mimic basic human senses (e.g. humans’ visual and auditory systems) [51], [52], [53] , [54]. Hence, we can use computers to assist human in their decision-making process and, by offering systems that consume accurate and re-trainable models, this process could evolve into a methodology filling the gap identified by Klir, 2017, as a pivotal contribution to his Uncertainty-Based Information Theory, level 4 [10, pp. 4–7].

Key Highlights of this Research

About Coffee Bean Roasting and its Impact on their Final Quality Grade

1. Statistical regression models were employed to study how the predictors of the roasted coffee bean parameters can explain their final quality grading. The regression

- models were not able of modeling the phenomenon. However, a machine learning model trained and validated using the same data, was able to modeled the prediction of the quality metric with robust accuracy (over 90%).
2. A deep neural network is capable of modeling domain expert knowledge by handling the underline phenomena between and quality indicators.
 3. Additional coffee data was collected at the University Lab (e.g., green and roasted bean color, moisture and kernel size, ambient temperature and roaster's gas and air flow levels, see Appendix 1F, 1G and, 1I). These additional parameters, combined with both roasting parameters and cupping paramaters (see Appendix 1E and, 1H) provide the foundation for the design of a Fuzzy Controller.
 4. The Fuzzy Controller will assist a domain expert, a human roast master, to roast coffee beans targeting their optimal quality while the bean is being roasted. This device's approach, theoretically, will roast coffee beans by looking at their final quality as an optimization problem to be solved, not as the coffee industry standard-protocol indicates (see Appendix 1B).

About Children Speech Intelligibility

1. First and foremost, the main hypothesis is confirmed that linguistic assessment generally reduces error and improves accuracy of intelligibility assessment.
2. It is possible to conclude that numeric procedure does not inflate estimates and the true estimate is actually captured by Linguistic terms which does not suppress/deflate the assessmen metric (i.e., DME).

3. Some sound subtypes are likely to be more accurately assessed using one methodology over another.
4. Statistically significant and positive coefficients are identified for Stop-affricate Consonants and Final Cluster-Final Singleton. So, the experimental results indicate that these word categories, or subtypes, are most effectively estimated using Linguistic scoring methods.
5. Linguistic assessment methods are not an improvement over assessment methods that combine Linguistic and Numeric, for any given single subtype
6. Results reinforce the idea of suggesting replacing numeric with linguistic holds less dispersion.
7. The similarity (or lack of difference) in the variability between the Numeric and Both treatments also drives the lower model fitness measure (as provided by the F-statistic)
8. Experienced Listeners reported higher dispersion (more error) than inexperienced.

Summary

This chapter provided an approach (de-erroring model) for quantifying judgment, a foundation for human knowledge representation and reasoning while making decisions in the presence of vagueness, imprecision, and uncertainty. Two highly complex systems whose behavior is not well understood have been given an approximated solution (hypothetically with less error), not only by modeling themselves but, by modeling their uncertainty as well by utilizing both linguistic scoring techniques and machine learning.

In this work, we have modeled how coffee roasting impacted the final quality of the beans and how measuring children's speech intelligibility as the child utters words by using direct magnitude estimation to improve speech intelligibility analysis. The aim is to integrate the naturalistic mechanisms of human decision making with machine learning capabilities as a useful framework with the hypothetical capacity of supporting several domains of the human decision-making process.

Essay 2: Computational Intelligence: Fuzzy Logic and Machine Learning Applied to Modelling Raw Beans Roasting and Judging (Cupping) of Specialty Coffees

Introduction

Motivation and Study Problem

Only for the past several years, researchers have noted the influence of roasting on coffee bean and cup quality [55] as an important factor driving the coffee industry. Currently, due to the heuristic nature of both the roasting and cupping (judging the final beans' quality), researchers are proposing, for example, to use fuzzy logic to model the human knowledge of coffee judges while performing a sensorial evaluation of coffee beans [32] [56] [57], and leaving behind the relationship between the roasting process and its impact into the final coffee bean quality. This research fills the gap in knowledge between the roasting process and its impact in the final coffee bean quality grading score.

Purpose and Research Approach

In this chapter, we take the novel approach of applying fuzzy logic combined with machine learning techniques, such as deep learning for modeling both, the roasting and its influence when determining the final quality of coffee beans through sensorial evaluation.

Moreover, a fuzzy controller is proposed to be integrated into a coffee roaster leveraging the human knowledge of the roast masters to drive temperature and airflow (output) from bean attributes such as moisture and color (input). Another fact, Machine Learning Algorithms (ML), a subset of artificial intelligence where computer algorithms are used to learn from data, are growing in popularity since the introduction of the Turing Test in 1950 to the learning capabilities of tracking human features exposed by Microsoft Kinect in 2010 [58].

For example, ML has proven to be useful in several industries ranging from pattern recognition in pharmaceutical research, automobile industry targeting self-driving cars, and identifying key attributes assessing the severity of heart failure [59]. Classification and clustering in data mining like big data algorithms currently in production systems covering a vast spectrum of services including the analysis of crime data [51]. However, ML potential is not fully explored in the coffee industry, in particular, specialty coffees. Moreover, in the food industry, for example, a machine vision system can facilitate the inspection of rice grains during processing for quality evaluation [60].

The Goals of the Study are to Support the Following Research Contribution

1. This research fills the gap in knowledge between the roasting process and its impact on the final coffee bean quality grading score
2. Aiming to design a fuzzy control system to model the heuristic element of the roast masters' domain knowledge, presenting the roasting process as a possible answer to an optimization problem targeting the best possible quality of the coffee beans
3. We proposed the unifications of roasting and cupping by combining the expert's knowledge from each of these domains

Organization of this Work

This document is divided as follows, in Section II a hypothetical statement is made through the lenses of ML, Section III includes a literature review, and Section IV presents the landscape of the research work covering coffee roasting, inferring the judging of final quality grading, a fuzzy control system design aiming to leverage the heuristic element of the roast masters' domain knowledge. Section V holds preliminary results along with the collaborative efforts needed for accomplishing this research. Section VI includes the conclusions and future work expected to be done to fully take advantage of the models presented in this research.

Research Questions

Can statistical analyses (i.e., regression analysis) explain the relationship between the process of coffee roasting and its impact on the final quality of the coffee beans?

Can machine learning techniques (i.e., deep learning) model the relationship between the process of coffee roasting and its impact on the final quality of the coffee beans, even without explaining the phenomena?

Research Hypothesis

The use of ML in this research is mainly to address the following hypothetical statement: “There is a relationship between the final quality-grading of a specialty coffee, determined by a judge through the sensorial evaluation of its attributes (e.g., fragrance/aroma, acidity, flavor, etc.) and the underlined parameters (i.e., charge and discharge temperatures, exhaust temperature, first and second pops' time and temperature, moisture, the color of the roast bean and so forth) measurable during the roasting process.”

Limitations

This research aims to study specialty coffees (coffees of high-quality) and, it departs from the standard protocols established by the Specialty Coffee Association (SCA), former Specialty Coffee Association of America (SCAA), see Appendix 2A, 2B, 2C and 2D). Additionally, the proposed design of a hardware device, a fuzzy roaster controller will also target specialty coffees and, in order to prototype this device, a collaborative effort is needed with other departments and institutions (i.e., electronic engineering, agriculture, etc.). For example, a formal invitation to collaborate with the Institute of Technology of Veracruz, Mexico, offers their coffee fields, years of experience studying coffee crops, their varieties and the process of coffee roasting; they have successfully prototyped and developed their coffee roaster using an industrial controller (see Appendix 2J). Furthermore, they have experimented using fuzzy logic to control a drum-based coffee roaster flame level (temperature) (see Appendix 2J and, 2K).

Key Terms

Coffee Cupping (judging coffee quality by an expert), Coffee Roasting (cooking the coffee beans in order to make them consumable by humans), Coffee Attribute (e.g., Aroma, Fragrance, Aftertaste, Body, Acidity, etc.). Specialty Coffees (i.e., coffees of high-quality), Specialty Coffee Association (i.e., organization that created the standards for the coffee industry). Regression analysis (i.e., a reliable method of identifying which variables have an impact on a topic of interest). Principal component analysis (PCA) (i.e., the technique used to emphasize variation and finds strong patterns in data). Deep Learning, a Machine Learning (ML) technique, a subset of Artificial Intelligence (AI).

Literature Review

This chapter aims to use statistical analyses and machine learning for modeling the specialty coffee bean roasting process and the final quality grading as the industry-standard indicates [42] [61]. The design and development of a fuzzy controller is proposed because the heuristic elements that coffee domain experts put into the process, make it practically impossible to model their knowledge.

A novel work from Godoy et al. [62] offers evidence of feasibility pertaining to fuzzy logic applied at industrial roasters in temperature control. They based this work on the fact that a fuzzy controller works well for systems lacking adequate methodology for its control, including the heuristic approach taken by coffee experts as they operate the roasters, making it very hard or impossible to achieve a satisfactory way of deriving a mathematical model of the roasting process, not to mention the cupping process.

Beyond controlling the roaster temperature, as proposed by Godoy et al., a fuzzy controller could also assist with air-flow control and exhaust air, key drivers when roasting small batches of specialty coffees. Research has shown that fuzzy logic offers a novel approach when mathematical models are not available for processes where human experts do well at mastering, making a good case for fuzzy logic [63].

Furthermore, Nogueira et al. [64] have proposed two multi-scale methods based on Convolutional Neural Networks (Deep Learning) to identifying coffee crops from Remote Sensing Images (RSIs) providing feedback to the Brazilian Agriculture-Industry. Their research laid the foundation for extracted geolocation of events (burned forest, for example), productivity forecast, and crop recognition from these images. In addition, this favors the identification of crops, a key process when monitoring the land-use, helping to define new expansion strategies of the land or to

estimate the feasible production amount.

Experimental Methods / Research Landscape

The section presents the details of the landscape where this research has been focused in terms of modeling raw beans roasting and their cupping or judging (sensorial evaluation) of their final quality grade.

Being the case that both the roasting and cupping, the obvious relationship of roasting a coffee bean to get it ready to be cupped (judge) is heuristically known by the roast masters and the coffee judges (cuppers). In order to apply machine learning, meaningful data must be collected from both the roasting and the cupping process, see section below titled “Incorporating Machine Learning (Random Forest Regressor).”

This research includes a solution to gathering all the data needed for the tracking of any interrelationships, interdependencies and, co-relationships between the roasting process (getting raw coffee bean ready for human consumption) and the cupping process (determining the coffee quality grade).

On the other hand, this research will lead the way toward the design and development of a fuzzy controller to be integrated into a coffee roaster used to roast high-quality coffee beans while machine learning techniques are employed to predict the final quality grading of the roast coffee beans.

Figure 2-1 plots a coffee roast curve. Critical cut points along the roast curve may be important drivers of the cupping results. Roasting parameters, as well as ambient environmental parameters (e.g., humidity) hypothetically impact the roasting process, with unknown and unmeasured impacts on the results (i.e., cupping to determine quality) of the coffee.

Hypothesized Functional Form of Cupping Results Driven by Roast Profile and Bean

Parameters

Cupping Vector [Roast (1-N)] = $f[(\text{Roast Vector}) + (\text{Growth Parameters Vector}) + (\text{Innate Bean-Specific Parameters}) + (\text{Unexplained Stochastic Properties})]$

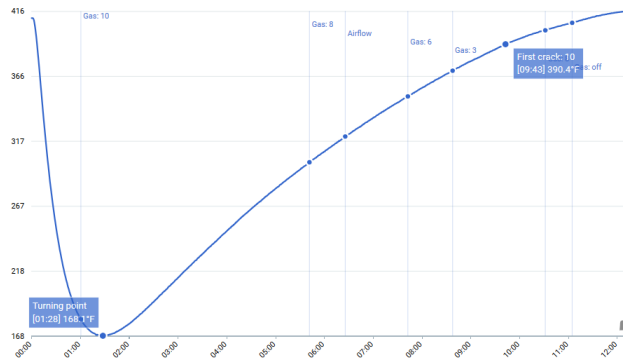


Figure 2-1. Coffee Roast Curve Plot, Y axis = temperature in Fahrenheit, X axis = time in Minutes

The following section aims toward the development and testing of the alternative measurement systems targeting the subsequent steps: tackling any interrelationships, interdependencies, and co-relationships between the roasting and the cupping processes. Figure 2-2 shows the four components of the Roast Profiler in charge of recording coffee bean batch profiles along with the origin of the coffee beans and their green parameters including variety, altitude, color, moisture, kernel size, etc.

The solution includes a client application interacting directly with the coffee roaster, a window-based application to manage collected data, a Restful API decoupling both the client

profiler, a windows-based application and, a relational database holding collected data.

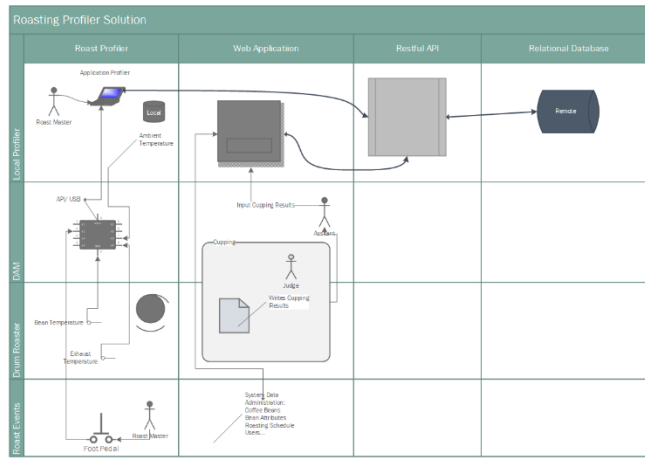


Figure 2-2. Roasting Profiler Solution

The roasting profiler solution, Figure 2-2, provides the UI, consisting of a windows-based client module, which interacts with a Restful Application Programming Interface (API). This API is based on the Model View Controller design pattern (MVC) [65], which achieves loose coupling between the constituent parts: the Model which represents the data, the View (the graphical display) and the Controller, in charge of managing the interactions (sending and receiving of messages) between the Model and the View.

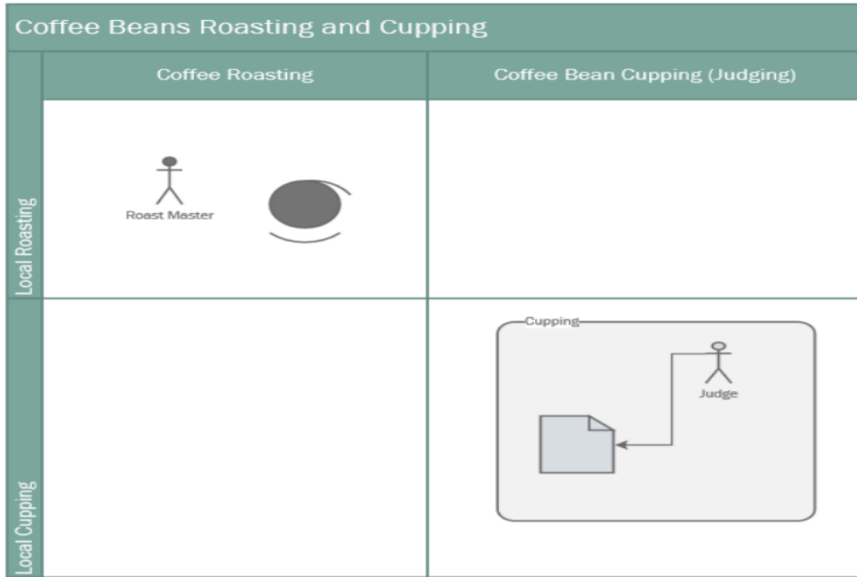


Figure 2-3. The Roasting and Cupping of Coffee Beans are two Separate Manual Processes

The SCA standard protocols described both, roasting and coffee cupping as two independent processes. In this chapter, we aim to study the impact of the roasting process on the output of the cupping process, i.e., the quality grade of specialty coffees. Figure 2-3 shows how the specialty coffee industry sees the roasting and judging processes. However, this research intends to unify these two processes by combining the expert’s knowledge from each of these domains. Before combining the data collected while studying these two industrial processes; in the next sections we present both, the roasting and the cupping processes in detail.

Roasting Specialty Coffee Beans

The chemical process by which aromatics, acids and other flavor components are either created, balanced or altered in the way that should augment the flavor, acidity, aftertaste, and body of the coffee as desired by the roast master (the person who roasts the beans) is called “Coffee Roasting” [61].

Furthermore, the dark color of the coffee is directly relative to the caramelization of the sucrose in the coffee. In the coffee industry, it is widely known that sucrose is a core determinant of both sweetness and bitterness. As a carbohydrate that makes up roughly fifty percent of the coffee's total dry weight by composition, sucrose has a direct impact on the class of important byproducts created during roasting are those of organic acids.

During roasting, proteins combine with carbohydrates in what is perhaps the most important reaction for all thermally processed foods – the Maillard Reaction. This set of reactions, discovered by a French chemist in 1910, is what is largely responsible for transforming the mere handful of compounds found in green coffee to the complex matrix that coffee is today [39]. This involves the evaporation of the bean’s moisture and reaction products from non-enzymatic and pyrolytic reactions evolve. Concluding, the most decisive parameter controlling the overall development of the coffee bean is the roasting temperature [55].

Inferring Specialty Coffee Bean’s Quality Grading

In this chapter, we study the process of judging specialty coffees as indicated in the Specialty Coffee Association (SCA) Cupping Form [42], derived from their Cupping Standard Protocol.

Figure 2-4 shows an excerpt from the latest SCA’ cupping form (see Appendix 2A).

Roast Level of sample	Score: <input type="text"/>	Score: <input type="text"/>	Score: <input type="text"/>
	Fragrance/Aroma	Flavor	Acidity
	6 7 8 9 10	6 7 8 9 10	6 7 8 9 10
	Dry	Qualities: Break	Aftertaste
	6 7 8 9 10	6 7 8 9 10	6 7 8 9 10
			Intensity
			High
			Low

Figure 2-4. SCA Cupping Form’s Strip, few Attributes of a Sample (Coffee Bean)

The Cupping Form (see Appendix 2A) provides means of recording eleven important attributes of the coffee bean as described in Table 2-1: Fragrance/Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Clean Cup, Sweetness, Overall, and to record Defects as well. These individual attributes are evaluated based on the scores shown in Table 2-2.

Additionally, the SCA protocol instructs the coffee judge (i.e., cupper) to rate coffee samples' quality grade using a numeric scale shown in Table 2-3. (see Appendix 2B, 2C and, 2D). In the current cupping process, crisp numeric values are used to represent coffee bean attributes perceived by the cupper; for example, Aroma equals 7.5, Acidity equals 7, and so forth.

Table 2-1. Coffee Bean Attributes as Described by the SCA. Seven attributes are classified as fuzzy attributes (*italics*), four attributes are crisp attributes (*normal*)

ID	Attribute: Description
01	<i>Fragrance:</i> The aromatic aspects include Fragrance (defined as the smell of the ground coffee when still dry) and Aroma (the smell of the coffee when infused with hot water).
02	<i>Flavor:</i> Represents the coffee's principal character, the "mid-range" notes, in between the first impressions given by the coffee's first aroma and acidity to its final aftertaste.
03	<i>Aftertaste:</i> Defined as the length of positive flavor (taste and aroma) qualities emanating from the back of the palate and remaining after the coffee is expectorated or swallowed.
04	<i>Acidity:</i> Is often described as "brightness" when favorable or "sour" when unfavorable.
05	<i>Body:</i> The quality of Body is based upon the tactile feeling of the liquid in the mouth, especially as perceived between the tongue and roof of the mouth.
06	Uniformity: Refers to consistency of flavor of the different cups of the sample tasted.
07	<i>Balance:</i> How all the various aspects of Flavor, Aftertaste, Acidity and Body of the sample work together and complement or contrast to each other is Balance.
08	Cleancup: Refers to a lack of interfering negative impressions from first ingestion to final aftertaste, a "transparency" of cup.
09	Sweetness: Refers to a pleasing fullness of flavor as well as any obvious sweetness and its perception is the result of the presence of certain carbohydrates.
10	<i>Overall:</i> The "overall" scoring aspect is meant to reflect the holistically integrated rating of the sample as perceived by the individual panelist.
11	Defects: Are negative or poor flavors that detract from the quality of the coffee.

Filling the SCA Cupping Form requires the following: first, the judge senses the coffee bean attributes. Second, he or she perceives the level of intensity of the sensed attribute and writes down a numerical score for the different attributes in the form. The cupper follows the following protocol's steps in filling the SCA Form: first, the coffee's Fragrance/Aroma is evaluated. Second, the Flavor, Aftertaste, Acidity, Body and Balance are evaluated. Third, Sweetness, Uniformity and Cleanliness including Overall score are decided on by the cupper. Finally, the cupper determines the sample quality score based on all of the combined attributes and deducts any faults (defects).

Table 2-2. SCA Coffee Attributes Quality Scale

Good	Very Good	Excellent	Outstanding
6.0	7.0	8.0	9.0
6.25	7.25	8.25	9.25
6.50	7.50	8.50	9.50
6.75	7.75	8.75	9.75

Table 2-3. SCA Coffee Final Grading Quality Scale

Good	Very Good	Excellent	Outstanding
$\geq 60 < 70$	$\geq 70 < 80$	$\geq 80 < 90$	≥ 90

A fuzzy expert system has been designed and developed to capture and preserve the irreplaceable human expertise of the coffee judges, see Figure 2-5.

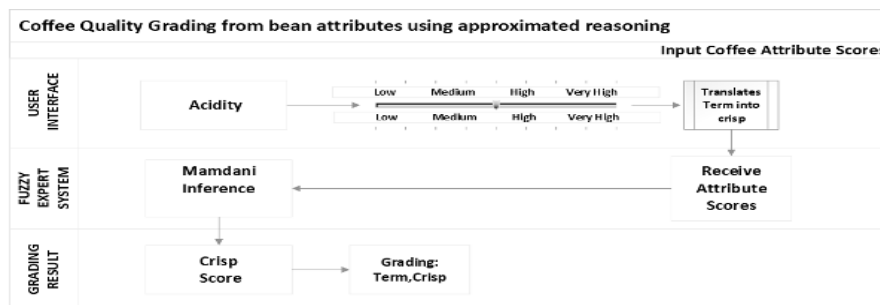


Figure 2-5. Coffee Bean Quality Grading Using Linguistic Terms for Attribute Scoring through the User Interface

The fuzzy expert system promotes the replacement of numerical values to express the individual attributes' scores with a process based on this selection of linguistic terms, Figure 1-5 shows how the attribute Acidity could be represented by any of the linguistic terms “Low, Medium, High or Very High.” However, the final grading of the coffee bean quality is not the arithmetic addition of the individual scores (as the SCA protocol indicates), it is the result of an approximated reasoning underlined by an appropriate Mamdani fuzzy engine, see Figure 2-5 - 6. Hence, offering a seemly level of abstraction between numbers and scores.

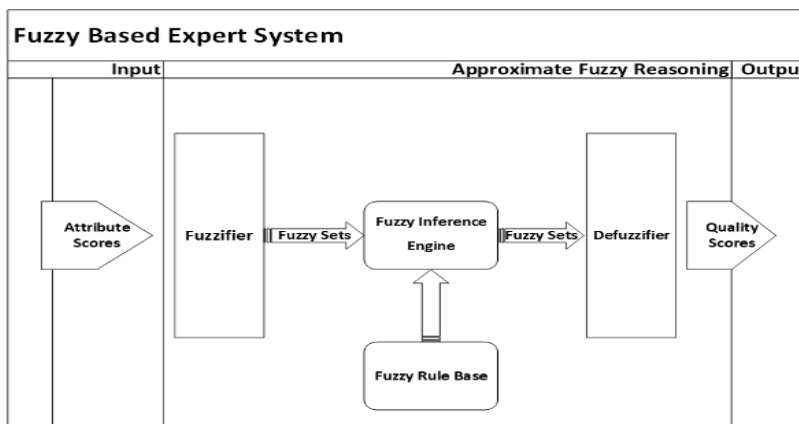


Figure 2-6 Fuzzy Expert System Architecture

Roaster Fuzzy Controller

Control systems are the greatest success of fuzzy logic. In areas where classical control knowledge has fallen short, fuzzy logic has proven to be effective and has become a very strong player as part of the industry standards [63, p. 437]. Figure 2-5, depicts a block diagram of a simple fuzzy control system proposed in [63, p. 442]. In this diagram, the steps (assumptions) included in Table 2-4 below, where *Plant*, is the physical system under control, it represents the roasting process which output (mainly heat and air-flow), at the same time, is determined by input signals coming from the coffee bean color and exhaust air moisture sensors.

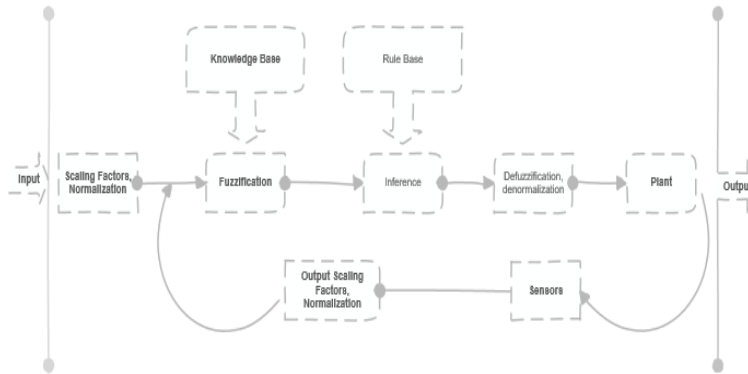


Figure 2-7. A simple Fuzzy Logic Control System Diagram [16, p. 442]

Moreover, the proposed fuzzy controller, by design will handle the bean temperature through gas flow, the exhaust air and, the air flow while monitoring the coffee bean’s color and the air (ambient) moisture. All of this with the awareness of other parameters like charge temperature, roaster temperature when the coffee beans fell into it, the first and possible the second cracks (pops), and so forth. Table 2-4 below renders a set of assumptions embraced by fuzzy controller designers. These assumptions were identified in [63, p. 441].

Table 2-4. Assumptions in a Fuzzy Control System Design [16, p. 441]

#	Assumption
01	<i>The plant is observable and controllable: state, input, and output variables are usually available for observation and measurement or computation</i>
02	<i>There exist a body of knowledge comprising a set of linguistic rules, engineering common sense, intuition, or a set of input-output measurements data from which rules can be extracted.</i>
03	<i>A Solution exists</i>
04	<i>The control engineer is looking from a “good enough” solution, not necessarily the optimum one</i>
05	<i>The controller will be designed within an acceptable range of precision</i>
06	<i>The problem of stability and optimality are not addressed explicitly; such issues are still open problems in fuzzy controller’s design</i>

After ascertaining the assumption listed in Table 2-4, the designer embraces a set of steps listed in Table 2-5.

Table 2-5. Steps in Designing a Fuzzy Control System

#	Steps
01	<i>Identify the variables (input, states, and outputs) of the plant</i>
02	<i>Partition the universe or discourse or the interval spanned by each variable into a number of fuzzy subsets, assigning each a linguistic label (subsets include all the elements in the universe of discourse)</i>
03	<i>Assign or determine a membership function for each fuzzy subset</i>
04	<i>Assign the fuzzy relationships between the inputs' states' fuzzy subset on the one hand and the outputs' fuzzy subsets on the other hand, thus forming the rule-based</i>
05	<i>Choose appropriate scaling factors for the input variables in order to normalize the variables to the [0,1] interval</i>
06	Fuzzify the inputs to the controller
07	Use fuzzy approximate reasoning to infer the output contributed from each rule
08	Aggregate the fuzzy outputs recommended by each rule
09	Apply defuzzification to form a crisp output

The core of a fuzzy controller design is the concept of its decision surface, a time-varying nonlinear surface, which encompasses the behavior of the controller's dynamics [63, p. 440]. Moreover, to approximate and build the control surface of a fuzzy controller, knowledge representation of a set of fuzzy conditional statements are paramount. This is a known fact, which makes fuzzy control design free from the paradigm of classical control linear system design. Fuzzy rule-based control systems are universal nonlinear function approximators which in essence, could be approximated to any desired precision [63, pp. 440-441].

Incorporating Machine Learning (Random Forest Regressor)

The goal of the cupping process is to allow the cupper (i.e., judge) to determine the final grading of the coffee bean quality as shown in Table 2-3. Any of these four-possible quality gradings, for example "Excellent" simply represents a final classification of the evaluated coffee bean. Hence, establishing the final quality grade of a coffee is a classification problem with four possible values: Good, Very Good, Excellent or Outstanding; the learning problem is to map the

predictor parameters from the roast profile with their quality grading scores, Table 2-6.

Table 2-6. Seven Predictors and the Output Parameter (Score)

ID	Attribute: Description
01	Initial Temperature: the temperature at which the roaster (machine) is pre-heated before the roasting process starts.
02	Minimum Temperature: value of the minimum recorded temperature during the roasting process.
03	Minimum Seconds: recorded time in seconds when the minimum recorded temperature was observed.
04	Maximum Temperature: value of the maximum recorded temperature during the roasting process.
05	Maximum Seconds: recorded time in seconds when the maximum recorded temperature was observed.
06	Final Temperature: value of the recorded temperature when the roasting process ended.
07	Seconds (roast time): total elapsed time of the roasting process.
08	Score: quality grade given to the coffee by the coffee judge after evaluating the coffee individual attributes

The final score, the label, representing the quality of the bean will be based on the values shown in Table 2-3.

Table 2-7. Several Parameters Combined from Green and Roast Data

ID	Attribute
01	Country of Origin
02	Region of Origin
03	Farm Altitude
04	Variety
05	Process
06	Green Color
07	Green Size
08	Roast Color
09	Roast Size
10	Air flow curve

Table 2-8. Sample Roast Data

COUNTRY	INITIALTEMP	MINTEMP	MINSECS	MAXTEMP	MAXSECS	FINALTEMP	SECONDS	SCORE
COLOMBIA	383	176	28	406	546	406	547	89
COLOMBIA	399	178	32	401	592	401	604	86.46
COLOMBIA	320	147	28	401	672	401	677	85.36
COLOMBIA	367	172	30	399	590	399	590	85.14
COLOMBIA	392	176	27	405	609	405	610	84
COSTA RICA	369	172	28	397	553	397	566	83.36
COSTA-RICA	361	183	31	396	608	396	619	84.92
COSTA-RICA	372	176	29	392	612	392	613	84.71
COSTA-RICA	372	178	26	401	564	399	567	84
DOMINICAN-R	374	176	28	396	577	396	582	83.68
EL-SALVADOR	381	181	33	410	675	288	687	85.71
EL-SALVADOR	379	174	31	397	621	397	627	84.92
EL-SALVADOR	392	178	29	410	588	410	594	84.46
EL-SALVADOR	390	181	31	401	678	401	681	82.67
ETHIOPIA	367	167	29	396	616	394	622	86.79
ETHIOPIA	376	180	30	394	602	394	605	85.04
ETHIOPIA	367	176	30	396	544	396	550	85
ETHIOPIA	378	174	30	396	578	396	581	84.46
GUATEMALA	385	172	28	396	642	396	652	83.79
GUATEMALA	385	171	29	410	679	410	679	82.85
HONDURAS	405.8	163.2	89	422.6	735	407	789	85

Table 2-9. Sample Data, Combining Origin and Roast

Country	Region	Altitude	Variety	Process	GreenColor* ^L	GreenColor* ^a	GreenColor* ^b	RoastColor* ^L	RoastColor* ^a	RoastColor* ^b	GreenBeansize-[µm] at C3 - 10%	GreenBeansize-[µm] at C3 - 50%	GreenBeansize-[µm] at C3 - 90%	GreenSize-Mn3(x) [µm]	GreenMoisture%	RoastMoisture%	RoastBeansize-[µm] at C3 - 10%	RoastBeansize-[µm] at C3 - 50%	RoastBeansize-[µm] at C3 - 90%	RoastSize-Mn3(x) [µm]
COLOMBIA	Planadas	1900	Caturra	Washed	44.18	0.64	12.22	29.73	4.73	6.01	4492	5899	6915	5807	8.51	0.37	5320	6883	8103	6805
COLOMBIA	Planadas	1900	Caturra	Washed	44.18	0.64	12.22	31.15	4.64	6.12	4492	5899	6915	5807	8.51	0.3	5335	6908	8085	6813
COLOMBIA	Planadas	1900	Caturra	Washed	44.18	0.64	12.22	30.09	4.47	6.27	4492	5899	6915	5807	8.51	0.28	5328	6908	8130	6808
COLOMBIA	Planadas	1900	Caturra	Washed	44.18	0.64	12.22	28.97	4.59	5.63	4492	5899	6915	5807	8.51	0.25	5336	6850	8061	6741
COLOMBIA	Planadas	1900	Caturra	Washed	44.18	0.64	12.22	29.45	4.81	5.81	4492	5899	6915	5807	8.51	0.22	5301	6878	8121	6793
COLOMBIA	Planadas	1900	Caturra	Washed	44.18	0.64	12.22	30.37	5.06	6.62	4492	5899	6915	5807	8.51	0.2	5386	6961	8209	6870
COLOMBIA	Planadas	1900	Caturra	Washed	44.18	0.64	12.22	28.86	4.63	5.76	4492	5899	6915	5807	8.51	0.23	5303	6869	8129	6786
COLOMBIA	Planadas	1900	Caturra	Washed	44.18	0.64	12.22	30.01	4.51	5.98	4492	5899	6915	5807	8.51	0.2	5309	6882	8101	6791
COLOMBIA	Planadas	1900	Caturra	Washed	44.18	0.64	12.22	28.16	4.18	4.77	4492	5899	6915	5807	8.51	0.29	5296	6888	8152	6798
COLOMBIA	Planadas	1900	Caturra	Washed	44.18	0.64	12.22	30.54	5	6.74	4492	5899	6915	5807	8.51	0.28	5311	6887	8156	6806
COLOMBIA	Planadas	1900	Caturra	Washed	44.18	0.64	12.22	29.74	4.74	5.99	4492	5899	6915	5807	8.51	0.25	5304	6882	8167	6805
HONDURAS	Intibuca	1600	Catuai	Washed	44.57	1.49	13.36	28.09	4	4.73	4444	6267	7302	6064	8.42	0.29	5381	7267	8487	7099
HONDURAS	Intibuca	1600	Catuai	Washed	44.57	1.49	13.36	27.28	4.6	4.48	4444	6267	7302	6064	8.42	0.38	5381	7267	8487	7099
HONDURAS	Intibuca	1600	Catuai	Washed	44.57	1.49	13.36	28.77	4.73	5.49	4444	6267	7302	6064	8.42	0.47	5353	7191	8452	7040
HONDURAS	Intibuca	1600	Catuai	Washed	44.57	1.49	13.36	26.37	3.85	3.82	4444	6267	7302	6064	8.42	0.33	5394	7191	8392	7050
HONDURAS	Intibuca	1600	Catuai	Washed	44.57	1.49	13.36	29.71	4.95	5.65	4444	6267	7302	6064	8.42	0.43	5348	7155	8468	7026
HONDURAS	Intibuca	1600	Catuai	Washed	44.57	1.49	13.36	28.35	4.39	4.36	4444	6267	7302	6064	8.42	0.26	5328	7161	8421	7004
HONDURAS	Intibuca	1600	Catuai	Washed	44.57	1.49	13.36	20.27	4.89	6.57	4444	6267	7302	6064	8.42	0.36	5354	7101	8349	6992

Discussion of Preliminary Results

Linear Regression Models, Looking for Linearity in the Data

In order to have a sort of benchmark, a foundation to work with, some work has been done in order to evaluate if a linear model could explain the correlation among the predictor parameters and the quality grading score shown in Table 2-6. We developed three statistical models trained

using data as shown in Table 2-8, only a subset of the data shown in Table 2-9. One model including all countries and two models, one with only Colombian coffees and another with Honduran coffees. Figure 2-8 shows the distribution of scores per country.

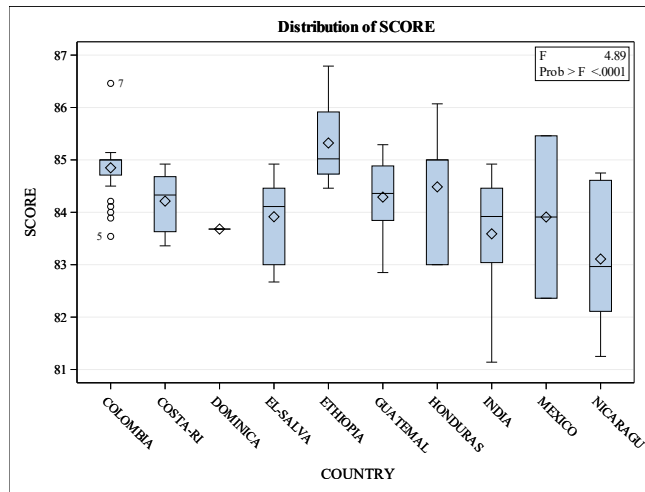


Figure 2-8. Dependent Variable Score distributed by Country

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	18.79427	2.68490	3.36	0.0024
Error	143	114.40776	0.80005		
Corrected Total	150	133.20203			

Figure 2-9. ANOVA Shows Statistical Significance of the Regression Model (All Countries)

Root MSE	0.89446	R-Square	0.1411
Dependent Mean	84.40722	Adj R-Sq	0.0991
Coeff Var	1.05969		

Figure 2-10. Model Could Explain 14 Percent of the Variability of the Data (All Countries)

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	90.16770	7.73991	11.65	<.0001
INITIALTEMP	1	-0.00132	0.00949	-0.14	0.8898
MINTEMP	1	-0.06441	0.02467	-2.61	0.0100
MINSECS	1	-0.00404	0.00941	-0.43	0.6682
MAXTEMP	1	0.01727	0.01488	1.16	0.2477
MAXSECS	1	0.00011692	0.00029946	0.39	0.6968

Figure 2-11. Only the Intercept Shows Statistical Significance (All Countries)

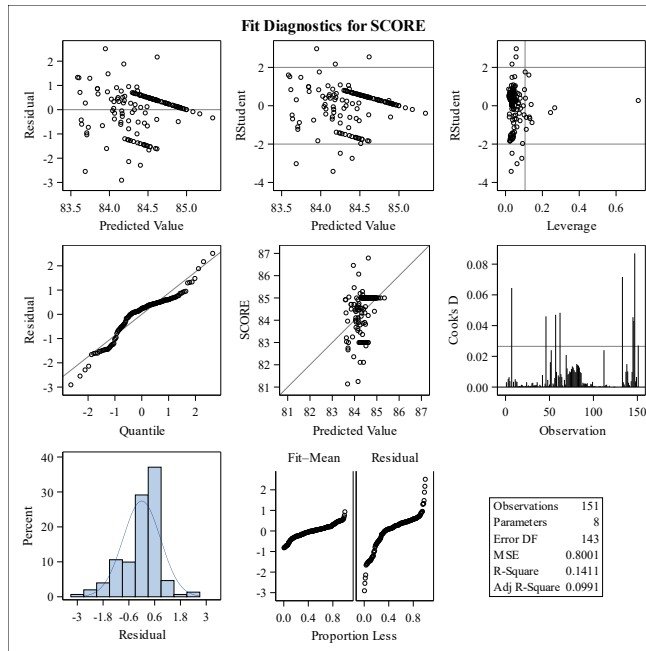


Figure 2-12. Fit Supporting Model Assumptions (All Countries)

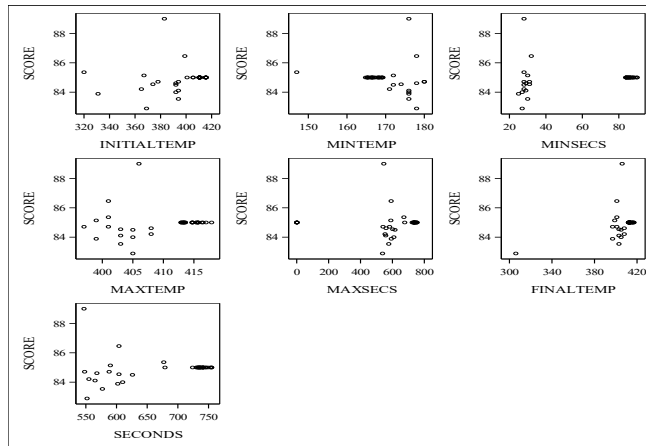


Figure 2-13. Scatter Plots (Colombia)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	6.18413	0.88345	1.24	0.3116
Error	32	22.83335	0.71354		
Corrected Total	39	29.01748			

Figure 2-14. Model is not Significant (Colombia)

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	108.19457	32.05082	3.38	0.0019
INITIALTEMP	1	0.01118	0.01480	0.76	0.4556
MINTEMP	1	-0.07030	0.05267	-1.33	0.1914
MINSECS	1	0.02327	0.02629	0.89	0.3827
MAXTEMP	1	-0.04176	0.06744	-0.62	0.5401
MAXSECS	1	-0.00006442	0.00054622	-0.12	0.9069
FINALTEMP	1	0.02046	0.00953	2.15	0.0394
SECONDS	1	-0.01244	0.00861	-1.44	0.1584

Figure 2-15. Individual Predictors Statistical Significance (Colombia)

Root MSE	0.84471	R-Square	0.2131
Dependent Mean	84.91675	Adj R-Sq	0.0410
Coeff Var	0.99476		

Figure 2-16. Model Only Explains 21 Percent of Variability (Colombia)

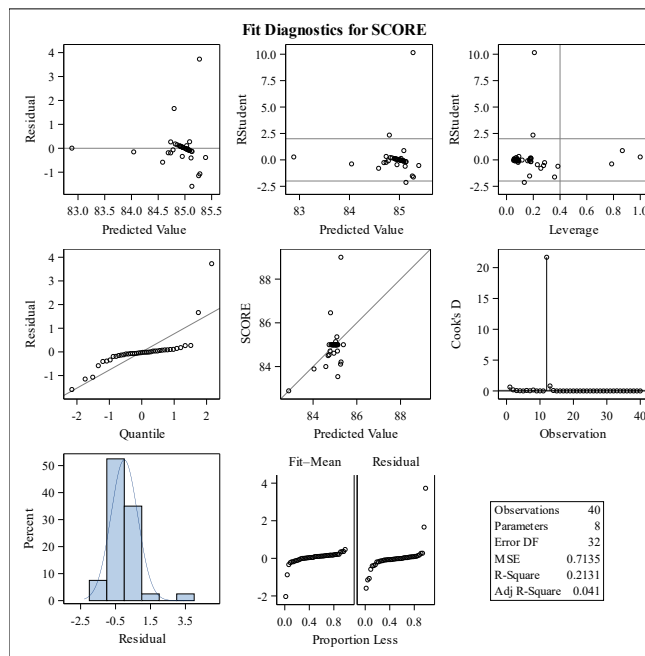


Figure 2-17. Fit Supporting Model Assumptions (Colombia)

The model assumptions are confirmed, residuals follow a normal distribution, Figure 2-17. The variance from the histogram Figure 2-13, is large and scatter plots look clustered. Some potential outliers were identified and removed. Scatterplot and regression analysis were done

without the identified potential outliers, the p-value for ANOVA is greater than 0.05, linear model does not fit, Figure 2-14 and Figure 2-15. Parameter Estimates all are not significant except for the intercept. R-square is 0.2131, Figure 2-16. Results do not justify a need for a modified model.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	2.98201	0.42600	2.53	0.0370
Error	29	4.88947	0.16860		
Corrected Total	36	7.87148			

Figure 2-18. Model is not Significant After Removing Outliers (Colombia)

Removing the outliers improved r-square slightly (it is now 0.3788, Figure 2-18, before was 0.2131, Figure 2-16). Parameter estimates aside from the intercept are still not significant. ANOVA Table, p-value 0.0370 (greater than 0.05) indicates that the data does not follow a linear model. QQ plot and histograms are improved by removing the outliers. Data points are more random around the residual plots. We did not perform the Box Cox transformation because the equality of variances assumption was not violated.

Null Hypothesis, $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$

Alternative Hypothesis, H_a : not all β_i are zero

Under H_0

H_0 , the statistic F_0 follows an F-distribution with 7 and 29 degrees of freedom. Because $P\text{-value} = 0.0370, P(F_{7,29} \geq 2.53) > 0.001$, we should Accept H_0 , this means that there is not a linear association between the score and the seven predictor variables. Moreover, the list below holding parameter estimates does. Figure 2-25, does not include any predictor variable's P-value indicating significance correlation. In other words, none of the predictors variables seems to have a linear relationship with the score.

The above facts do not justify further analysis including sequential procedures like Forward Selection, Backward Propagation nor Stepwise Regression because these methods are based on the P-value.

We can calculate the F_0 (the test statistic) for $\alpha = 0.05$, meaning with 95% of confidence:

$$H_0: \beta_1 = \dots = \beta_p = 0$$

$$H_a: \text{not all } \beta_k \text{ (i = 1; : : : ; p) equal zero}$$

$$F_0 = \frac{2.98201 / 7}{4.88947 / (37-7-1)} = 2.52669 \text{ (df 7,29), } P = 2.35 > 0.05$$

Reject H_0 if $F_0 \geq F_p; n-p-1; \alpha$. In this case we reject the null hypothesis (H_0) because $F_0 > P$.

On the other hand, to avoid multicollinearity we will not consider interaction terms. These terms will include variables which are part of the model.

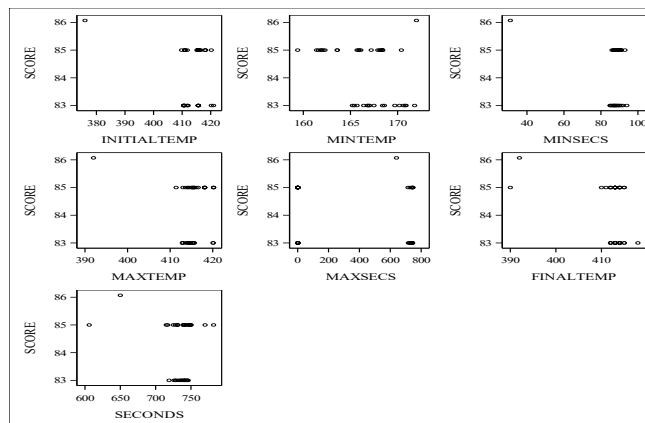


Figure 2-19. Scatter Plots (Honduras)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	13.07842	1.86835	2.06	0.0769
Error	32	28.95686	0.90490		
Corrected Total	39	42.03528			

Figure 2-20. Model is not Significant (Honduras)

Root MSE	0.95126	R-Square	0.3111
Dependent Mean	84.17675	Adj R-Sq	0.1604
Coeff Var	1.13008		

Figure 2-21. Model Only Explains 31 Percent of Variability (Honduras)

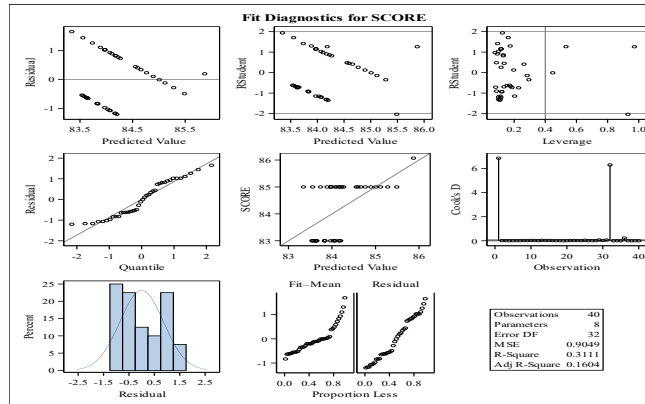


Figure 2-22. Fit Supporting Model Assumptions (Honduras)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	10.82083	1.54583	1.74	0.1360
Error	31	27.53814	0.88833		
Corrected Total	38	38.35897			

Figure 2-23. Model is not Significant After Removing Outliers (Honduras)

Root MSE	0.94251	R-Square	0.2821
Dependent Mean	84.12821	Adj R-Sq	0.1200
Coeff Var	1.12033		

Figure 2-24. Model Only Explains 28 Percent of Variability (Honduras)

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	82.43533	44.76041	1.84	0.0751
INITIALTEMP	1	-0.03096	0.11965	-0.26	0.7976
MINTEMP	1	-0.10100	0.06995	-1.44	0.1588
MINSECS	1	0.06489	0.07798	0.83	0.4116
MAXTEMP	1	0.12228	0.11641	1.05	0.3016
MAXSECS	1	-0.00001217	0.00090793	-0.01	0.9894
FINALTEMP	1	-0.07778	0.09042	-0.86	0.3963
SECONDS	1	0.00931	0.01255	0.74	0.4638

Figure 2-25. Individual Predictors Statistical Significance (Honduras)

$H_0: \beta_1=\beta_2=\beta_3=\beta_4=\beta_5=\beta_6=\beta_7= 0$, H_a : not all β_i are zero

Under H_0 , the statistic F_0 follows an F-distribution with 7 and 32 degrees of freedom. Because $P\text{-value} = 0.2581$, $P(F_{10,28} \geq 1.34) > 0.001$, we should Accept H_0 , this means that there is not a linear association between the score and the seven predictor variables. Moreover, the list below holding Parameter Estimates does not include any P-value indicating significance. In other words, none of the predictors variables seems to have a linear relationship with the score. This fact does not justify further analysis including sequential procedures like Forward Selection, Backward Propagation nor Stepwise Regression because these methods are based on the P-value.

Robustness test of the Machine Learning Model -- Multivariate Principal Component Analysis (PCA)

The linear regression models did not provide enough explanation of the variability in the data. First of all, not all assumptions for applying linear regression to the data were met (e.g., the data is not normally distributed, there is not a linear relationship amongst the variables, nor little multicollinearity).

Table 2-10. Parameters Correlation Table

	INITIALTEMP	MINTEMP	MINSECS	MAXTEMP	MAXSECS	FINALTEMP	SECONDS
INITIALTEMP	1.0000						
MINTEMP	0.2166	1.0000					
MINSECS	0.5800	0.0165	1.0000				
MAXTEMP	0.8117	0.1593	0.4553	1.0000			
MAXSECS	0.2441	0.2141	0.4025	0.3018	1.0000		
FINALTEMP	0.6684	0.1233	0.3550	0.8246	0.2559	1.0000	
SECONDS	0.1038	0.1693	0.3592	0.1330	0.9294	0.0504	1.0000

Table 2-10. shows high correlation (i.e., multicollinearity) amongst some variables (bolded values in table). For example, the highest correlation (0.9294) is observed between the variable *MAXSECS* (time in seconds when the roast profile reaches it maximum temperature) and

SECONDS (total roasting time in seconds). Another very high correlation is reported with the value of 0.8246 between *MAXTEMP* (i.e., the maximum temperature value recorded in the roast profile for a particular batch of coffee) and, the *FINALTEMP* (coffee bean temperature recorded at the moment of the roast profile discharge). These high correlations suggest the use of a technique such as PCA. PCA is a method that maximizes the variance of the data dimension. Considering the multivariate analyses based on the true eigenvector-based methods, PCA is the simplest and best for this scenario [66]. PCA aims to reveal the unseen structure of the data by explaining its variability (i.e., variance) [66].

Table 2-11. Principal Component/Eigenvalues

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	3.26544	1.547290	0.4665	0.4665
Comp2	1.71815	0.733817	0.2455	0.7119
Comp3	0.984336	0.371823	0.1406	0.8526
Comp4	0.612513	0.363440	0.0875	0.9401
Comp5	0.249072	0.125685	0.0356	0.9756
Comp6	0.123387	0.076291	0.0176	0.9933
Comp7	0.0470956	-	0.0067	1.0000

Table 2-11. Holds the Eigenvalues for our data. These values account for the variance of the component (i.e., Comp1 = 3.26544). The components are listed from the highest variance down to the lowest. Moreover, the column titled “Proportion” gives the value of the percentage of the variance accounted for the component. Component one (Comp1) explains 46.65 percent of the variation of the data and, component two (Comp2) explains 24.55 percent. These two components combined explain the 71.19 percent of the variation of the data (See Appendix 2M).

For our study, we will pick only the components with Eigenvalues greater than one (e.g., Comp1, Comp2 and, Comp3) (See Appendix 2N and, Appendix 2O). The 71.19 value represents a considerable improvement of the linear regression models presented at Figures 2-15, 2-26, 2-20, 2-23, 2-24 and, 2-25 where the maximum percentage explaining the data was 28.

A Random Forest Regressor (RFR) was Used

A random forest is a classifier consisting of a collection of tree-structured classifiers $\{h(x, \Theta_k), k=1, \dots\}$ where the $\{\Theta_k\}$ represent independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x . RFR “they always converge so that overfitting is not a problem.” [67, p. 4].

For this RFR Model, the same data used for all countries, see Table 2-7. All fields are treated as continuous variables. Furthermore, we identified the dependent variable (i.e., SCORE), the value to be predicted. In this model, we set the training ratio with the value of 0.80 (80 percent of the data will be used to train the RFR y 20 percent used for testing). Root-mean-squared percent error is the metric we have used, Figure 2-26. The list of parameters used by the RFR model are shown in Figure 2-27 where the dependent variable is highlighted and, seven independent variables are shown (e.g., FINALTEMP, INITIALTEMP, etc.). Additionally, in order to prepare the data for the model to use the Root-mean-squared percent error, the SCORE’s values were transformed using the natural logarithmic function, see Figure 2-28.

```
In [8]: 1 def rmse(x,y): return math.sqrt(((x-y)**2).mean())

In [9]: 1 # RMSLE (root mean squared log error) between the actual and predicted scores.
2 # Therefore we take the log of the scores, so that RMSE will give us what we need
3 df_raw.SCORE = np.log(df_raw.SCORE)
4 df_raw.SCORE

Out[9]: 0      4.442651
1      4.430817
2      4.430817
3      4.442651
4      4.442651
...
1236   4.406719
1237   4.488636
1238   4.276666
1239   4.488636
1240   4.330733
Name: SCORE, Length: 1241, dtype: float64
```

Figure 2-26, Root-mean-squared Metric Used in the RFR Model

```
In [10]: 1 display_all(df_raw.isnull().sum().sort_index()/len(df_raw))

FINALTEMP      0.0
INITIALTEMP     0.0
MAXSECS        0.0
MAXTEMP        0.0
MINSECS        0.0
MINTEMP        0.0
SCORE          0.0
SECONDS        0.0
dtype: float64
```

Figure 2-27, List of Parameters Available for the Model

	INITIALTEMP	MINTEMP	MINSECS	MAXTEMP	MAXSECS	FINALTEMP	SECONDS	SCORE
0	320	147	28	401	672	401	677	1.491251
1	365	171	28	408	554	408	555	1.488584
2	394	176	29	403	558	403	565	1.488584
3	378	180	28	397	541	397	548	1.491251
4	394	180	31	401	582	401	588	1.491251

Figure 2-28, Excerpt of the Data fed Into the RFR Model

```
In [16]: 1 # n_estimators is the number of trees to be used in the forest.
2 # Since Random Forest is an ensemble method comprising of creating multiple decision trees,
3 # this parameter is used to control the number of trees to be used in the process.
4
5 number_of_estimators = 32
6 m = RandomForestRegressor(n_estimators=number_of_estimators, bootstrap=True, n_jobs=-1)
7 m.fit(X_train, y_train)

Out[16]: RandomForestRegressor(bootstrap=True, ccp_alpha=0.0, criterion='mse',
max_depth=None, max_features='auto', max_leaf_nodes=None,
max_samples=None, min_impurity_decrease=0.0,
min_impurity_split=None, min_samples_leaf=1,
min_samples_split=2, min_weight_fraction_leaf=0.0,
n_estimators=32, n_jobs=-1, oob_score=False,
random_state=None, verbose=0, warm_start=False)
```

Figure 2-29, Setting up the Random Forest Regressor Model

Table 2-12 shows a list of experiments designed for testing the impact of the PCA components with the highest eigenvalues (e.g., Comp1 and, Comp2) see Table 2-11. For example, the experiment for control (i.e., Control) includes all the variables, trial one excludes highest components from principal component Comp1 and, trial two excludes only highest component from principal component Comp2. All of this to assess the marginal contribution of each of the

three largest factors from the PCA components (e.g., INITIALTEMP, MAXTEMP and, FINALTEMP).

Table 2-12. Experiments for Testing the Impact of PCA Components in the Machine Learning Model

	Control	Trial one	Trial two	Trial three	Trial four	Trial five	Trial six
	Included	Included	Included	included	Included	Included	Included
INITIALTEMP	x		x		x		x
MINTEMP	x	x	x	x	x	x	x
MINSECS	x	x	x	x	x	x	x
MAXTEMP	x		x			x	x
MAXSECS	x	x			x	x	x
FINALTEMP	x		x		x	x	
SECONDS	x	x	x	x	x	x	x

The RFR Model ran for each of the experiments listed on Table 2-12 and, the results are shown on Table 2-13.

Table 2-13. Eigenvector Omitted by Components

Trial	Accuracy	Eigenvector Omitted from Comp1	Eigenvector Omitted from Comp2	Marginal RFR Model Accuracy Decrease for Change in Eigenvector
Control	0.85	0.0000	0.0000	0.0000
one	0.81	1.3580	-0.9098	0.0295
two	0.83	0.3592	0.5482	0.0557
three	0.81	1.7172	-0.3616	0.0233
four	0.83	0.4732	-0.2962	0.0423
five	0.82	0.4610	-0.2879	0.0651
Six	0.84	0.4246	-0.3257	0.0236

The eigenvectors omitted by Components shown in Table 2-13 (see Appendix 2P and, Appendix 2Q) are reflecting the marginal contribution of these PCA components (e.g., Comp1

and, Comp2) when the variables with the highest impact are removed from each of the experiments shown in Table 2-12, the column titled “Explained” holds the R squared (R^2) reported by the RFR Model when the trial was completed. The R^2 is the proportion of the variance in the dependent variable (i.e., SCORE) predictable from the independent variables included in each of the trials, see Table 2-12 and Figure 2-17.

The Table 2-13 column titled “Marginal RFR Model Accuracy Decrease for Change in Eigenvector” holds the decrease of the RFR Model per trial. For example, for trial one (i.e., 0.0295), this value results from:

Control Accuracy (85%) minus trial one accuracy (81%) divided by Comp1 eigenvalue (1.3580):

$$\frac{0.85 - 0.81}{1.3580} = \frac{0.04}{1.3580} = 0.0295$$

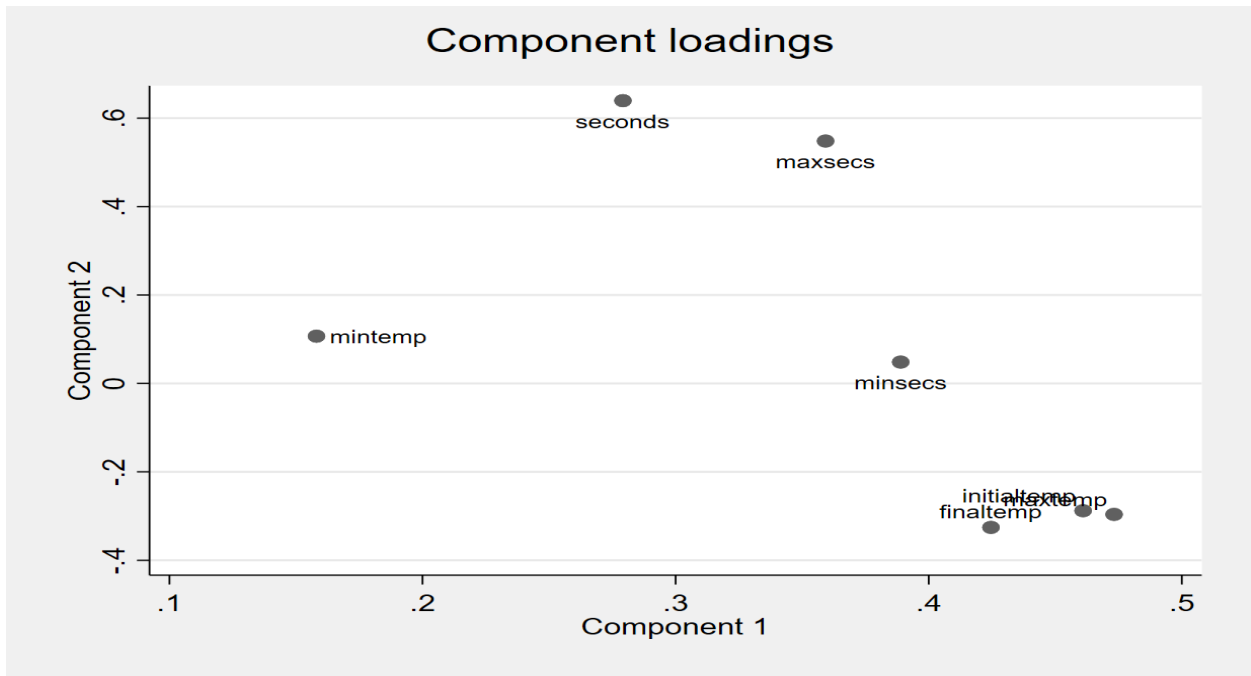


Figure 2-30, PCA Component Loading (Comp1 and, Comp2)

Furthermore, Figure 2-30 indicates that for PCA Comp1, the three variables with the highest impact (i.e., eigenvectors) in the RFR Model are initial temp (0.5431), max temp (0.5579) and, final temp (0.5304). Moreover, for the PCA Comp2, the variables with the highest impact are seconds (0.6936) and, MAXSECS (0.6543).

Moreover, in Table 2-13 the columns titled “Eigenvector Omitted from Comp1” and titled “Eigenvector Omitted from Comp2” reflect the eigenvector values that were omitted during the execution of a trial by the RFR Model. For example, no eigenvector values (from Comp1 nor Comp2) were omitted by the Control experiment. However, trial one has the following values (e.g., for Comp1 is 1.358 and for Comp2 is -0.9098). These values resulted for the following calculations:

Eigenvector Omitted from Comp1 (1.358) = Comp1 {Eigenvalue – Eigenvalues [initial temp + max temp + final temp]}

Eigenvector Omitted from Comp1 (1.358) = 3.26544 – (0.5431 + 0.5304 + 0.5579)

Eigenvector Omitted from Comp1 (1.358) = Comp2 {Eigenvalue – Eigenvalues [INITIALTEMP + MAXTEMP + FINALTEMP]}

Eigenvector Omitted from Comp1 (-0.9098) = 1.71815 – (- 0.0201 – 0.0212 – 0.0709)

Using the experimental procedure just identified to selectively omit explanatory variables from the RFR, we also identify strong consistency of the machine learning model in both directions, which provides an even stronger robustness check. With comparisons across trials, we derive a clear linear relationship that is negatively associated with the omission of larger eigenvector values, and positively associated with the omission of larger negative eigenvector values. Results suggest that RFR decreases its accuracy (i.e., explains less) as we omit more. For example, if we take away (high value characteristics such as initial temperature), it hurts the RFR

performance. The marginal decreasing accuracy, as shown on Table 2-13 for trial one, trial three and, trial five are 0.0295, 0.0233 and, 0.0651 being the initial temperature (i.e., INITIALTEMP) as the parameter with the highest impact.

Additionally, all the variables representing the maximum metrics (i.e., MAXSECS) are the least impactful (i.e., they caused the RFR Model to have a lower accuracy). The trials having the best results are trial one, and trial three (these trails omit MAXSECS).

Conclusions and Future Work

Results indicate that the RFR is capable of predicting the quality grading scores with accuracy over eighty percent, Table 2-12. The RFR does not place the SCORE's values far away from the quality grading ranges showed in Table 2-3 as the standard suggested. The goal is to make the system always produce reliable predictions of the coffee bean quality grade by keeping its accuracy under control, meaning within a reasonable range. This could be achieved by incorporating a fuzzy controller into the roasting process modeling the heuristic of the domain experts (roast masters) in charge of roasting the coffee beans considering both coffee bean origin, green parameters and, roasted parameters obtained during the roasting process itself (see Appendix 2F, 2G) and those parameters measured at the lab after the coffees have been roasted (see Appendix 2I).

A future work that could potentially enhance the accuracy of predicting the quality of the coffee bean while it is being roasted. This process will employ a type-2 fuzzy set (2FS), an extension of the currently used. Research has shown that 2FS breeds high accuracy when dealing with uncertainties [26]. The rationale is that the currently used fuzzy engine computes the input values of the attributes' individual scores, in a similar way as a probability distribution function computes the mean, which measure-of-dispersion is captured by the variance [26, p. 643]. By using 2FS, the

fuzzy model could rely on a more suitable measure-of-dispersion, one capable of handling the uncertainties inherent to the linguistic terms used by the domain experts' roasting and judging the coffee beans. In other words, fuzzy logic type I addresses data imprecision while 2FS handles the expert imprecision. As shown in Table 2-6, the RFR Model used seven predictors, all from the roasting process, used to predict the final quality grading score (final result from the non-deterministic roasting process), leaving out of the model several other measurable parameters (see Appendix 2G, 2I).

Additionally, further work can apply ML to understanding the relationship between the way coffee beans have been roasted and their final quality grade, including several more coffee bean parameters (e.g., moisture, color, density, kernel size, and farm altitude). This could be evaluated running simulation tests to give and take away parameters as the PCA suggests; our experimental statistical analysis serves an important validation function by confirming that the selective withholding of statistically-influential component loadings from the machine learning model weakens the explanatory power of the machine learning model in the expected direction, for both positive and negative vector loadings.

Nonetheless, the coffee origin or green parameters as shown in Table 2-7, are yet to be understood considering that the coffee bean moisture is considerably reduced as well as the beans' kernel size, Table 2-7.

Furthermore, in order to select an optimal subset of these parameters, a genetic algorithm is proposed as it has been successfully employed by Santhanam, T. and Padmavathi, M.S. [68] with the goal to minimize the number of selected features that maximize the accuracy of the classification as suggested by Sohrabi and Tajik [69].

Moreover, modeling the final quality grading of the coffee bean and its relationship with

the way it has been roasted seems to be a good approach. Nonetheless, modeling the final consumer's acceptance (suggested by Professor Oscar Gonzalez Rios from TNM/Instituto Tecnológico de Veracruz) as they are the one paying for the final product, it is a more noble approach due to the fact that the coffee judge who determined the grading, is not the only one tasting the coffees.

Additionally, collaborative efforts at the Veracruz Institute of Technology, with other areas like the electronic department, industrial engineering, agriculture, biochemistry and experts on sensory evaluation of coffee beans are keen players on this research (see Appendix 2L).

Essay 3: The *IntelliTurk* Tool for Evaluating Speech Intelligibility in Children: Software Tool Introduction and Experimental Validation

Introduction

Motivation and Study Problem

Efficient, accurate, reliable, and valid measurements of speech intelligibility are a critical component of determining the severity of functional communication in everyday environments. Improvement of speech intelligibility is a primary aim in the remediation of speech disorders in children [70]. Although an established consensus exists amongst clinical speech pathologists supporting the importance of evaluating speech intelligibility, little agreement exists on how to measure it [71] [72]. Most clinicians rely on subjective impressions of intelligibility yielded from intuition or by extrapolating a level of intelligibility based on measures of speech severity [73].

Recently, more objective methods, like rating scales, have been used to measure speech intelligibility. Rating scales typically require a listener to mark a visual analog, such as an equal-appearing interval or a labeled ordinal scale [74]. Interval scale measures have faced scrutiny, however, as listeners are challenged to accurately and reliably divide the scale [75].

Purpose and Research Approach

Magnitude estimation has been used to resolve the issues regarding perceptual partitioning of speech intelligibility across a continuous scale [74] [76]. In this case, an anchor sample, termed the modulus, is provided as an example of a speaker with a determined level of intelligibility. The listener then rates impressions of the magnitude of difference between the modulus and the speech sample presented. This approach has yielded promising results, but the method is not without limitations.

Limitations of Existing Approaches

Listeners may be challenged to rate the magnitude of difference from the modulus (e.g., twice as intelligible, or three times less unintelligible), or they may revert to providing numeric values without estimating the magnitude of difference. Additionally, the severity of the modulus may introduce a listening bias that influences ratings of intelligibility. Some evidence supports the standardization of the modulus at a medial level of intelligibility for direct magnitude estimation experiments [77]. Magnitude estimation may be further influenced by the unit of measurement being observed, which may lack a definite unit of measurement or have numeric values that are visualized across a scale.

Listener characteristics also influence the evaluation of intelligibility. Listener level expertise with a particular type of speaker or speech disorder, as well as with the materials used to make assessments, affect their impressions of intelligibility [78] [79]. Moreover, it is well known that information (data) on any conceptualized problem has limited utility as a direct result of the uncertainty associated either with the experts or with the methodology and tools engaged during problem-solving tasks [80, pp. 245-246]. Additionally, information resulting from the reduction of uncertainty neglects the essence of what human communication and cognition are born from: the richness of human notions based on their surroundings, perceptions, previous experiences, and more. These thoughts, aligned with our research commitment to develop a specific mathematical theory that supports the modeling of our subject of research, are factually limited by the constraints of the theory itself [80, p. 246].

Research Approach

In this work, we use linguistic terms in combination with crisp (i.e., numeric values) terms to capture the elasticity of human perception. This is an initial step towards integrating Fuzzy Logic with empirical findings based on magnitude estimation. We look at Direct Magnitude Estimation (DME), as a linguistic variable and in a fuzzy set holding a group of terms [29, pp. 140-141] [81, p. 65]. The universe of discourse of the DME is derived from the numeric ranges (crisp values) from each of these linguistic terms, as “our natural language is the supreme expression of sets” [38, pp. 94-96]. These numeric ranges define the degree of membership of the individual values ranging from Very Difficult, Difficult, Medium, Easy, and Very Easy, as shown in Table 3-2. This scale will be utilized by experienced listeners to indicate the level of intelligibility of speech samples reviewed in this experiment.

The goal is to take advantage of the fact that fuzziness is independent of the DME measurement obtained by the user selection. Selection of the linguistic term by the user carries some uncertainty, and this uncertainty translates to imprecision (a property of the phenomena itself as another attribute, dimension, or variable that would be a great topic for future study). Hence, as a Proof of Concept (POC), in this work has been developed the *IntelliTurk*[®] application and associated framework. *IntelliTurk*[®] has been created to support researchers when designing experiments that target either inexperienced or experienced listeners who are ascertaining child speech intelligibility through direct magnitude estimation. It is predicted that the applied use of AI will increase and that new clinical tools will emerge that automate clinical processes like intelligibility assessment [82]. A shift in research inquiry has emerged as a result of improved speech recognition performance. As large datasets have become available, more advanced modeling methods based on deep learning has yielded increasingly intelligent solutions (that go

beyond regression models) when solving complex computational problems. However, further research is needed to develop useful applications of deep learning models for automatic speech intelligibility detection, specifically within the pediatric population, that capture both the abnormal speech variation and subjective ratings of assessors of speech intelligibility [82].

This study aims to enable researchers to customize the scope of their experiments based on (a) the target audience (i.e., experienced or inexperienced), (b) their preferred scoring method for collecting DME metrics (i.e., Numeric, Linguistic, and the combination of Numeric and Linguistic), and (c) the use of a default or a custom reference recording, as needed for listener training (as suggested by Stevens [78]). Additionally, researchers will have the option to run the experiments without the Amazon Mechanical Turk, and can instead run experiments in their lab or classrooms. In summary, this study will accomplish the following:

- Identify levels of speech intelligibility.
- Develop tools that effectively aid clinicians in the assessment of child intelligibility and use deep learning for computational speech assessment.
- Reduce ambiguity in analyzing levels of children's speech.
- Enhance the ability of speech pathologists to create experiments with the *IntelliTurk* framework.

Research Questions

Which scoring mechanism, Numeric, Linguistic, or Both (Numeric/Linguistic), results in less error when applied to collect DME metrics?

Which group of listeners, Experienced or Inexperienced, scores DME metrics with less variability?

Research Hypothesis

The null hypothesis being tested is: whether or not Linguistic assessment methodologies improve variability in intelligibility assessment. These child intelligibility experiments allow us to test the hypothesis that there is no statistical difference in DME values between Numeric, Linguistic, and Numeric/Linguistic user estimation approaches (see Tables 3-2 and 3-3).

Limitations

This chapter focuses on child intelligibility, and it is based on experiments targeting children. A limitation of this study is that it was performed on a group of children that were opportunistically sampled at a university testing lab. Thus, the results cannot be generalized to the entire population. Nevertheless, this research has fewer limitations as a control for the experiment for Direct Magnitude Estimation (DME), as the research participants were crowd-sourced based on entrance criterion from the Amazon Mechanical Turk.

Key Terms

Direct Magnitude Estimation (DME), Amazon Mechanical Turk (AMT), Speech Intelligibility, Numerical Scoring, Linguistic Scoring, External Stimulus, Machine Learning

Literature Review - Established Intelligibility Assessment Methodology

The way a human perceives and judges external stimuli has laid the foundation for the development of scaling mechanisms, such as the expression of perceived magnitude of stimuli, the ordinal discrimination judgments of stimuli, and the partition of the sensory continuum into equal

intervals. These methods are leveraged on “the basic assumption,” the idea that humans can correctly ascertain a situation’s intensity [83].

Researchers have been encouraged to look for a suitable mechanism to model domain-specific phenomena that are influenced by human’s subjective intuition, such as the expression of perceived magnitudes of stimuli. The expression of perceived magnitudes of stimuli is a psychophysical scaling method used to explore the relationship between human sensation and physical stimuli, and researchers have widely employed it since its introduction [78].

Research has shown that a human’s ability to judge duration, distance, and intensity is core for their brain to build mental representations, one such phenomenon being the modality-independent capacity of language [84]. Researchers in the field of magnitude estimation had employed a variety of dependent measures [85]. For example, past work has allowed subjects to adjust the loudness of a sound, brightness of a light, vibration frequency, numeric estimates, and line lengths, among others [84].

Since the early sixties, researchers have advocated for the exploration of emerging disciplines, such as information theory, space perception, and multidimensional scaling [85]. Additionally, they have conducted a set of pioneering studies in the fundamental areas of human perception that suggest alternative approaches for modeling the judgment of multidimensional stimuli. Their concerns hovered around the fact that from the perspective of stimuli, researchers must attempt to understand the effects of different attributes or dimensions and the combination of intensities inferred from them.

Moreover, the explanation of the extensive set of empirical findings on magnitude estimation, one of the oldest topics in psychology, remains central to researchers’ investigations. For example, a group of researchers has suggested a generic principled mechanism for perceptual

inference, a well-founded methodology under the general framework of Bayesian inference. These researchers successfully developed a modeling framework for distinguishing between shared and selective representations of magnitude in neuroimaging experiments, and showed how their association (i.e., brain activities) with cognitive distortions is noted in psychiatric disorders [86].

An interactive software labeled MEDAS, the Magnitude Estimation Data Analysis System, was developed by Sung et. al. in 1997 [83], to equip researchers with systematic and generalized methods for analyzing magnitude estimation data. MEDAS housed a set of procedures and deterministic rules designed to assist researchers during the analysis of collected magnitude estimation data. It allows the researchers to choose the appropriate data standardization method, mean deviation, min-max values, and the geometric mean method [83, p. 520]. Furthermore, MEDAS frames the rules and procedures among three different stages of the magnitude estimation data analysis (i.e., the analysis of the response space, cross-modality matching/merge, and data standardization) and for the analysis of two different types of scales. These scales, the unipolar and bipolar scales, were designed to support what is called “the response space.” Whereas the unipolar scale aims to allow an experiment to go directly to the stage of cross-modality, the bipolar scale favors the treatment of negative responses; when both scales are present, they should be combined to constitute one sensory continuum [83, p. 516]. However, to date, MEDAS does not support the design or gathering of experimental data. To the best of our knowledge, there is no tool currently available that supports both the designing and gathering of experimental data.

Furthermore, the internet was not as mature when MEDAS was developed as it is today. Currently, tools such as Amazon Mechanical Turk® (AMT), a web-based platform that supports crowdsourcing (a method of obtaining information through the online recruitment of many non-expert listeners), facilitate streamlining listener recruitment and speech intelligibility assessment

processes. The AMT offers a set of services, including compensation and data gathering, at a reduced cost and with high flexibility: “Crowdsourcing is a good way to break down a manual, time-consuming project into smaller, more manageable tasks to be completed by distributed workers over the Internet” [87].

Experigen [88] has recently been offered for download online and at no cost to facilitate the integration of linguistic experiments through AMT. This platform is a script-based experimental design tool that allows researchers to design experiments using a set of templates and a subset of libraries. To maximize the research experiment’s compatibility with the current array of web browsers, the Experigen support team suggests employing standard HTML, CSS (Cascade Stilling Sheets), or JavaScript depending on the targeted audience. However, this approach requires that researchers pay attention to these standards in addition to designing, running, and collecting the experiments’ data. Experigen comes with some documentation in the form of a checklist that walks researchers through how to download, install, set up, and run experiments. This approach clearly demands some knowledge of text editors, such as TextWrangler for OS X or Notepad++ for Windows-based computers. Nonetheless, the researcher needs basic knowledge of the source control public repository (GitHub) where Experigen source code is kept. At the time of drafting this paper, the latest comment of the Experigen source code, labeled “extra files for artificial language experiments,” was timestamped on September 18, 2014 [89].

These technical skill requirements limit those researchers without coding experience and burden them with acquiring personnel who possess this specialized skill set. Experigen creators acknowledge on their GitHub software page that the use of their platform requires webpage design

and deployment skills and R (R Core Team, 2013) statistical computing package knowledge. Further, it is stated that “It’s not for linguists who are not good with computers.”

Moreover, Kawahara conducted a detailed survey of psycholinguistic methodologies in phonological research [90]. In this work, Kawahara presents perspectives on how data are often collected, for example using fieldwork research and dictionary inquiries, and several resources are listed, including a comprehensive list of books, quantitative methods, and some software (such as the Praat, doing phonetics by computer), that feature a set of phonetic analyses, such as acoustic and perception. The R statistical package offers scripting functions that facilitate the automation of repetitive processes, such as resampling and data processing. Many resources designed for perceptual experiments have been deployed for research-specific projects that are limited to the aims of that specific study. Hence, to address the lack of currently available resources for the study of the perceptual phenomenon, the *IntelliTurk: Speech Intelligibility and Perception Platform for AMT*, Speights Atkins *et al.*, 2019 was created.

System Design

IntelliTurk has an architecture driven by decoupled modules, independent components that interact through message passing (i.e., experiment’s results), as shown in Figure 3-1. These messages bring the experiment’s target audience and the subject responses to the survey, the screening, and the study where the word rating process is performed.

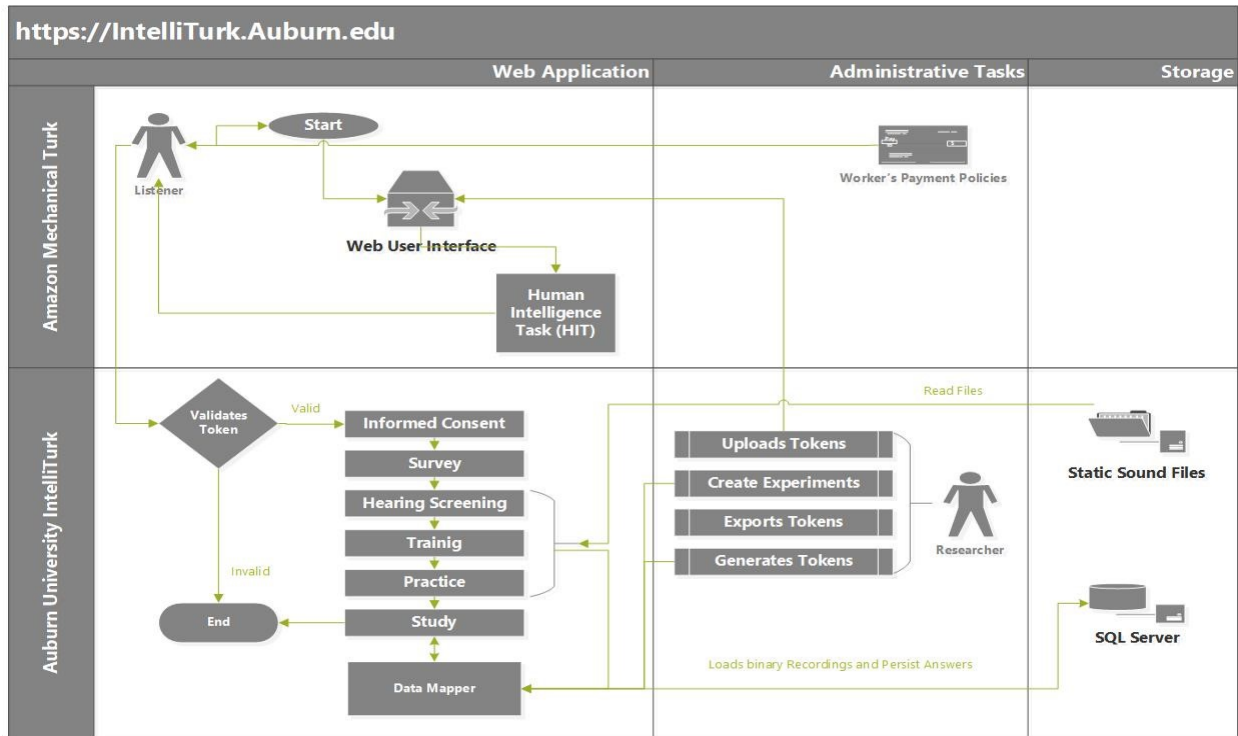


Figure 3-1. *IntelliTurk* detailed solution architecture diagram.

The *IntelliTurk* human-computer interaction's main element is a slider bar that facilitates the user's entry of a valid value by providing a range of values for selection. Figure 3-2 depicts how the *IntelliTurk* guides a subject to rate a heard word using the slider.

The screenshot shows the user interface for 'Sound 1 of 100'. It includes a 'Play Sound' button, a text input field for the user's response, and a 'Rate the word' section with 'Re-play Sound' and 'Play reference sound' buttons. A slider bar is used for rating, with a 'Value: 57.5 Medium' displayed. The slider has labels for 'Very Difficult', 'Difficult', 'Reference Sound Level: 50 (Medium)', 'Medium', 'Easy', and 'Very Easy'. At the bottom, there are 'Next >' and 'End Participation' buttons, along with an 'AWS Worker Token'.

Figure 3-2. User intelligibility assessment response input.

Moreover, *IntelliTurk* facilitates the design of experiments to control for listener experience, including inexperienced listeners and experienced listeners (domain experts). Settings allow specifications for targeting the behavior of the slider, either with linguistic terms only, both linguistic terms and numeric values, or numeric values (crisp values) only. Figure 3-3 shows the experimental design administration page.

Figure 3-3. *IntelliTurk* experiment design administration page.

***IntelliTurk* Solution Architecture Design**

The *IntelliTurk* web application was developed using Microsoft .Net Technologies to provide a user-friendly interface for researchers without any scripting experience. The name was inspired by the project's name Intelligibility Study and AMT. The name *IntelliTurk* also reflects the integration between our web application and AMT. In the web-based application, administrators are provided with secured admin pages where they can create experiments and

generate tokens. These tokens allow for the integration of the *Intelliturk* application with AMT by exporting experiment-specific tokens into CSV files. The process of generating and uploading tokens is shown in Figure 3-3. AMT enlists workers, or internet users, to complete jobs referred to as Human Intelligence Tasks (HITs). The CSV files are uploaded into the HITs to be associated with a worker eventually. When a worker decides to conduct the listening experiment, they click on the provided URL and are directed to *intelliturk.auburn.edu*. A query string (i.e., the portion of the URL carrying back-end data) parameter containing a unique token for that specific worker will be added. Table 3-1 shows an excerpt of tokens, confirmation numbers, tokens' statuses, and some details as to whether a token has been used or not based on the worker's interaction with *Intelliturk*.

Table 3-1. Example List of Tokens and Confirmation Numbers Managed by *IntelliTurk*

Experiment Identifier	Token	Confirmation Number	Is Used	Tracker Status
1	3e048bc1-a684-4137-95bf-ab54a45c03e1	590008288	1	Worker could not fix sound problem
2	48dd6fbe-db68-4f1c-8eb1-e0703b1950be	28492377	1	Worker Completed the Experiment
2	48dd6fbe-db68-4f1c-8eb1-e0703b1950be	369688797	1	Worker Ended Participation
3	508d01d8-61be-41a0-9251-c475a1c2b1f3	786500977	1	Worker is not qualified (Survey Failed)
4	508d01d8-61be-41a0-9251-c475a1c2b1f3	605443273	1	Worker Completed the Experiment

Note: Example includes statuses.

When the worker (listener) lands at the default page of the study, *IntelliTurk* must validate the incoming token before they can start the survey. If the token has not been used, when its status value (Is Used) equals zero (0) and its tracker status value (Tracker Status) is empty, the users are presented with web pages as depicted in Figure 3-1. Otherwise, a page indicating an invalid token is rendered.

At the end of the study or when a listener stops their participation (through cancellation or abandonment, or the system has identified that they are not qualified to move forward), listeners

must provide a confirmation number (i.e., an unique, internally-generated number paired with the tokens used by AMT) at the AMT window. The *IntelliTurk* administrators or information technology leads use this confirmation number to identify the listener's completed tasks and proceed with compensation. Listener's responses in specific sections within the *IntelliTurk* are recorded in a database. Figure 3-2 shows one of the 100 sounds to be heard by the subject.

The user interface (UI) is rendered to the subjects to survey and validate their participation in the experiment. The subject is presented with a button to play the sound and a textbox to type what the subject heard (see Figure 3-2). As subjects navigate the study's webpages, *Intelliturk* constructs messages containing their responses regarding the pre-recorded speech samples and what they had identified. Constructed messages are sent to a database through the decoupled database library across the mapper. The mapper (i.e., in charge of translating the messages coming from the UI, into the corresponding entities in the database), is a library capable of exposing, partially part of the data stored in the database to the user, keeping both the UI and database completely decoupled.

Additionally, administrative pages (shown in Figure 3-3) are only accessible by authorized researchers and database administrators. These secured pages connect to the database's library, thereby isolating the database from the UI.

Nonetheless, the application works directly with AMT's remote workers. However, *IntelliTurk* also supports conducting experiments at the researcher's lab by allowing subjects to be initiated with a confirmation number, as seen in Figure 3-4.

Intelligibility Study

Welcome to the IntelliTurk - Intelligibility Study

About the Study

We are conducting an academic study about child speech intelligibility. We need listeners to rate and describe what they hear child speakers say. The study will begin with a brief consent, questionnaire, hearing test, training and practice questions, followed by a 100-item study. It is anticipated that the study will take at most 90 minutes to complete. **In order to compensate you for your participation, we will provide you with a unique confirmation number.**

Your work will be tracked using a controlled number which allows us to determine which items have completed. If for some reason, during the experiment, you experience technical difficulties, extraneous circumstances or lose internet connection before you complete the study, use the provided confirmation number to be compensated for the work you have completed and/or contact us at mis0096@auburn.edu.

If you have a confirmation number

Please enter your confirmation number below:

Please validate your Number

Figure 3-4. The *IntelliTurk* supports researchers in engaging subjects in their lab by using confirmation numbers.

Data Preparation and Analysis

The *IntelliTurk* platform allows for experiments that target both experienced and inexperienced listeners tasked with rating speech stimuli. In our POC test case, we applied the DME measurement tool as the method for rating the intelligibility of children’s speech. The DME tool can capture ratings under three conditions: numeric, linguistic, or a combination of both. These proposed conditions are covered by setting the experiment’s slider scope. Table 3-2 shows *IntelliTurk* slider linguistic terms and their overlap with numeric ranges.

Table 3-2. *IntelliTurk* Slider Linguistic Terms, Numeric Ranges, and their Overlap

Term	Very Difficult	Difficult	Medium	Easy	Very Easy
Ranges	10 to 15.6	26.9 to 38.00	49.3 to 60.4	71.7 to 82.8	94.1 to 100
Overlap	15.7 to 26.8	38.1 to 49.2	60.5 to 71.6	82.9 to 94.00	

The child intelligibility experiments allow us to test whether or not linguistic assessment methodologies improve the error in intelligibility assessment. In other words, the null hypothesis

being tested is: there is no statistical difference in DME values between Numeric, Linguistic, and Both (Numeric/Linguistic) user estimation approaches.

$$H_0 : \sigma_{DME_{Linguistic}} = \sigma_{DME_{Numeric}} = \sigma_{DME_{Both}} \quad (1)$$

Data were prepared for econometric modeling targeting a model either within or between variation (Figure 3-3). Table 3-4 shows a breakdown representing the various data-points being captured throughout the experiments.

Table 3-3. Study Targeting Hypotheses Covering Three Treatments and Two Subjects Types

Treatment	Ho (Null Hypotheses)	Ha (Alternative Hypotheses)
Numeric Only	(μ I) DME = (μ E) DME	(μ I) DME != (μ E) DME
Linguistic Only	(μ I) DME = (μ E) DME	(μ I) DME != (μ E) DME
Both	(μ I) DME = (μ E) DME	(μ I) DME != (μ E) DME
(Numeric/Linguistic)		
Numeric Only	(σ I) DME = (σ E) DME	(σ I) DME != (σ E) DME
Linguistic Only	(σ I) DME = (σ E) DME	(σ I) DME != (σ E) DME
Both	(σ I) DME = (σ E) DME	(σ I) DME != (σ E) DME
(Numeric/Linguistic)		

Note: I=Inexperienced, E=Experienced, μ =mean, σ =Standard Deviation

Table 3-4. Breakdown of the List of Fields (Data-Points) Captured During Experiments

Field (data-point)	Brief description
<i>Subject</i>	String Token used as an interface with AMT
<i>Question</i>	Question Number (value between 1 and 100)
<i>Target</i>	The word contained in the sound file, expected to be heard by the subject (listener)
<i>DME</i>	Direct Magnitude Estimation Value (coming from <i>IntelliTurk</i> UI Sliders)
	When Numeric is equal to 1 or when Both is equal to 1
<i>SubjectID</i>	Unique Confirmation Number used as a Subject Identifier
<i>Inexperienced</i>	Categorical Value (1 indicates an inexperienced listener)
<i>Experienced</i>	Categorical Value (1 indicates an Experienced listener)
<i>Numeric</i>	Categorical Value (1 indicates <i>IntelliTurk</i> UI Sliders Show Only Numbers)
<i>Linguistic</i>	Categorical Value (1 indicates <i>IntelliTurk</i> UI Sliders Show Only Linguistic Terms)
<i>Both</i>	Categorical Value (1 indicates <i>IntelliTurk</i> UI Sliders Show Both Numbers and Linguistics Terms)

When the slider’s value falls into any of the covered overlap’s range listed in Table 3-2, a random number between 1 and 10 is generated. Thus, if this random number is less than or equal to 5, the left linguistic term is chosen; otherwise, the right term is selected. For example, when the slider value is 40 (overlap in the range 38.1 to 49.2) and the random number is 6, the term “Medium” is selected. However, if the random value is 5, the term “Difficult” is selected.

Validation Study Research Method

Speech Samples

Speech samples were retrieved from an Auburn University Institutional Review Board (IRB)-approved repository of speech samples of children both with and without disorders (see Appendix 3A). Speech samples were recorded in a quiet room. The environmental noise level was determined to be below 40 dBA SPL for each recording session [91]. Samples were recorded at a 44K sampling rate with 24-bit depth using handheld H6N recorders with cardioid XLR MOVO LV402 microphones. Speech samples consisted of words within eight phonetic contrast categories: (1) stop-fricatives, (2) stop-affricates, (3) final cluster-final singletons, (4) fricative-affricates, (5) alveolar-palatals, (6) front-back vowels, (7) high-low vowels, and (8) initial cluster-initial singletons [77]. These phonetic contrasts have been affiliated with reduced intelligibility in children with phonological-based disorders [71]. The speech sound subtypes used in the phonetic contrast categories are defined in Table 3-4.

Table 3-5. List of Speech Sound Subtypes

Category	Definition	Group
Stop	Produced with complete closure at a specific point in the vocal tract.	Manner of Articulation
Fricative	Produced when active and passive articulators approximate each other so closely that the escaping air causes an audible friction	Manner of Articulation

Affricate	Produced in two phases: the first phase includes complete closure formed between the active and passive articulator followed by a slow-release resulting in the friction of the sound (a combination of a stop and fricative)	Manner of Articulation
Alveolar-palatal consonants	Produced in the anterior cavity of the oral cavity at the place of the alveolar ridge vs. produced with constriction at or near the palate	Place of Articulation
Front-back vowels	Vowels produced in the anterior versus the posterior portion of the oral cavity	Place of Articulation
High-low vowels	Produced with the tongue position elevated toward the palate or lowered away from the hard palate.	Place of Articulation
Initial cluster-Initial singleton	A sequence of consonants preceding the vowel versus a single produced consonant preceding the vowel.	Syllable structure process
Final cluster-final singleton	A sequence of consonants that occur after the vowel versus a single produced consonant that follows a vowel.	Syllable structure process

One-hundred words belonging to nine categories (i.e., non-contrast, NC; stop-fricative, S-F; stop-affricate, S-A; final cluster-final singleton, FC-FS; fricative-affricate, F-A; alveolar-palatal, A-P; front-back vowels, F-BV; high-low vowels, H-LV; and initial cluster-initial singleton, IC-IS) were presented to each listener. The first category of non-contrast words includes nine words selected from the Clinical Assessment of Articulation and Phonology Second Edition [92]. These words comprise randomly ordered speech sounds from each of the eight categories (e.g., leaf, cheese, dog, swing). Words in the remaining eight categories are considered contrast pairs, meaning that paired words in a category differ by only one target speech sound (e.g., “chip” versus “ship”), creating a contrastive pair. The contrast categories represent common errors made by young children and children with speech disorders that affect intelligibility [93].

Children Speakers (Research Population)

Nine child speakers with varying levels of word production accuracy due to age and speech health status (non-disordered vs. disordered speech) were selected from the database. The sample

included 5 males and 4 females. The ages of children ranged from 3 y, 4 m to 5 y, 5 m. Each child was assessed for the presence of a speech sound disorder using the Diagnostic Evaluation of Articulation and Phonology [94]. Scores are based on a scale of 10 and a standard deviation (SD) of 3. A score of 7, one SD below the mean, was utilized as the criterion for determining the presence of a speech sound disorder. Six children exhibited non-disordered speech, while three children exhibited disordered speech. All child speakers demonstrated the following characteristics: (1) normal bilateral hearing at 20 dB for 0.5 kHz, 1 kHz, 2 kHz, and 4 kHz; (2) American English as their primary language; and (3) the ability to orally communicate at least one-word utterances.

Child speakers were categorized into three groups, high, medium, and low, according to whole word production accuracy measured by the Proportion of Whole-Word Correctness (PWC) [95]. A PWC score is the proportion of accurately produced words in a speech sample. PWC was calculated for each speaker from orthographic transcripts recorded by two trained graduate students. Disagreements in the transcripts were resolved by consensus. Three additional trained graduate students independently coded each word, transcribed as “0” for incorrect transcriptions that did not match the intended word and “1” for correct transcriptions that matched the intended word. From this code, we obtained a percentage of the total number of words identified correctly by the speaker when compared to all the words spoken in the dataset. Discrepancies in coding were resolved by consensus. Speakers were classified into three groups according to the PWC score of high (>85%), mid (50% to 84%), and low (<50%). The same stimuli such as pre-recorded words were presented to participants within the speech intelligibility measurement experiment.

Preparation of Speech Samples: Materials and Procedure

One-hundred experimental listening stimuli represented words from each phonetic category and the high, mid, and low speaker groups. Each stimulus list began with the same eight sound files consisting of single-syllable word productions. The remaining files from the phonetic contrast groups were randomized to control for list order and speaker effects. The same words were presented to each listener in the same order.

Experimental Design

We evaluated our hypothesis using a randomized controlled experiment consisting of three treatment groups. The treatment groups (previously identified in Table 3-3) are referred to as *Numeric*, *Linguistic*, and *Both (Numeric/Linguistic)*. In the first and second treatments, subjects only observed Numeric and Linguistic DME rating input entries, respectively. In the third treatment, subjects observed both Numeric and Linguistic DME rating entries. By operationalizing all three treatments by groups and comparing them in an econometric model (described below), we can functionally determine DME estimation differences for the same sound recordings separately by rating input entries. Simply put, the experiment provides the ability to fully control for exogenous differences that might otherwise explain DME estimation differences aside from estimation methodology. In a controlled randomized experiment, any remaining differences are due to systematic influences associated with treatment effects, in this case, the intelligibility assessment methodologies of Linguistic, Numeric, or Both.

Table 3-6. Experiments Targeting Listeners Through the Amazon Mechanical Turk (AMT)

Using Auburn University's Experimental Design Platform, *IntelliTurk*

Listener's Experience Level	Group	Treatment	Number of Subjects
-----------------------------	-------	-----------	--------------------

Inexperienced	1	Numeric Only	10
Inexperienced	1	Linguistic Only	10
Inexperienced	1	Both (Numeric/Linguistic)	10
Experienced	2	Numeric Only	10
Experienced	2	Linguistic Only	10
Experienced	2	Both (Numeric/Linguistic)	10

Two groups of experiments were conducted, one on a group targeting inexperienced listeners (Group 1), and the other on a group of confirmed experienced listeners (Group 2). The subject population consisted of adult listeners with no more than incidental exposure to child speech. They were recruited for this study through the AMT crowdsourcing platform, which allows for data collection from a diverse population. AMT workers selected the HITs titled “Child Speech Intelligibility Study.” After accessing the HIT, AMT workers gained access to the research experiment through an embedded *IntelliTurk*[®] link. Those who selected the link and agreed to complete the HIT were assigned specific token numbers and associated confirmation codes for de-identified administrator task review and compensation. Workers blindly self-selected a HIT in the AMT from one of three visual conditions: (1) only numeric values, (2) only linguistic values (ranging from “Very Easy” to “Very Difficult” to understand), or (3) both numeric and linguistic values (see Tables 3-2 and 3-6).

Participants qualified for participation if they were from the United States, were speakers of an American English dialect, had no hearing impairment, and did not regularly interact with children at home or work between the ages of 2 and 7 years.

At the beginning of the study, each listener was instructed to complete the experiment in a quiet place while using headphones with the volume set to a comfortable listening level. Listeners were required to verify that their computers and headphones were functioning properly before advancing. Speech recognition ability was screened within the *IntelliTurk*[®] web application using the Word Intelligibility Picture Identification Test [96] [97]. This word

recognition task was initially designed for children but has been used to evaluate listener performance in adults for experimental purposes [98].

Additionally, listeners completed training and practice items for task conditioning before beginning the experiment. Listeners were instructed to type the word they heard and then to rate the intelligibility of the speech according to a referent recording determined to be in the middle intelligibility range by three experienced speech-language pathologists. Ratings were made using the DME tool's dynamic sliding bar. The point selected on the sliding bar yielded a quantitative DME score and, when specified, a linguistic term of intelligibility for each condition. The study items were transcribed and rated according to the same reference recording introduced in the training.

The empirical analysis of experimental results consists of formalized hypothesis testing and regression estimation.

Results

For the inexperienced group (Group 1), 49 AMT workers responded to the HIT. Thirteen workers were excluded from the study due to not meeting the inclusion criteria, and two ended participation. A total of 34 listeners were eligible for inclusion in the study.

In Group 2, out of 34 experienced listeners who responded to the HIT, 32 completed the study. Two listeners ended the study before completion.

Hypothesis Tests

To evaluate and measure treatment differences between the three experimental treatment groups, we formally tested the differences using formalized statistical tests. First, it was necessary to test whether or not there existed differences in the mean DME values by treatment. Our

hypothesis is predominantly about variance (i.e., error) in the DME estimates rather than mean differences. In other words, we are testing if linguistic assessment methodologies improve the error in intelligibility assessments, but we have not specified any hypotheses regarding bias (defined as an error in a consistent direction). In this case, bias would be measured as a formalized difference in the treatment mean.

Hence, we tested for mean differences to rule out bias and focus on the more important issue at hand—error. The mean and SD of observed DME estimates by treatment are presented in Tables 3-7 and 3-8. From a comparison of the means, the treatment groups do not have different DME estimates. However, observations indicate that the differences are more generally associated with the variance of those estimates. Additionally, initial observations were supplemented by conducting formalized Wilcoxon tests (i.e., non-parametric Mann-Whitney tests) of treatment equality. These tests all failed to reject the null hypothesis.

Table 3-7. DME Estimate Descriptive Statistics Inexperienced Listeners (Group 1)

Treatment	Mean	St. Dev.	Obs.
<i>Numeric</i>	54.141	27.957	1,508
<i>Linguistic</i>	54.528	24.789	1,200
<i>Both</i>	57.096	26.366	1,201

Table 3-8. DME Estimate Descriptive Statistics Experienced Listeners (Group 2)

Treatment	Mean	St. Dev.	Obs.
<i>Numeric</i>	47.889	31.495	1,000
<i>Linguistic</i>	49.354	29.755	1,200
<i>Both</i>	50.822	29.076	800

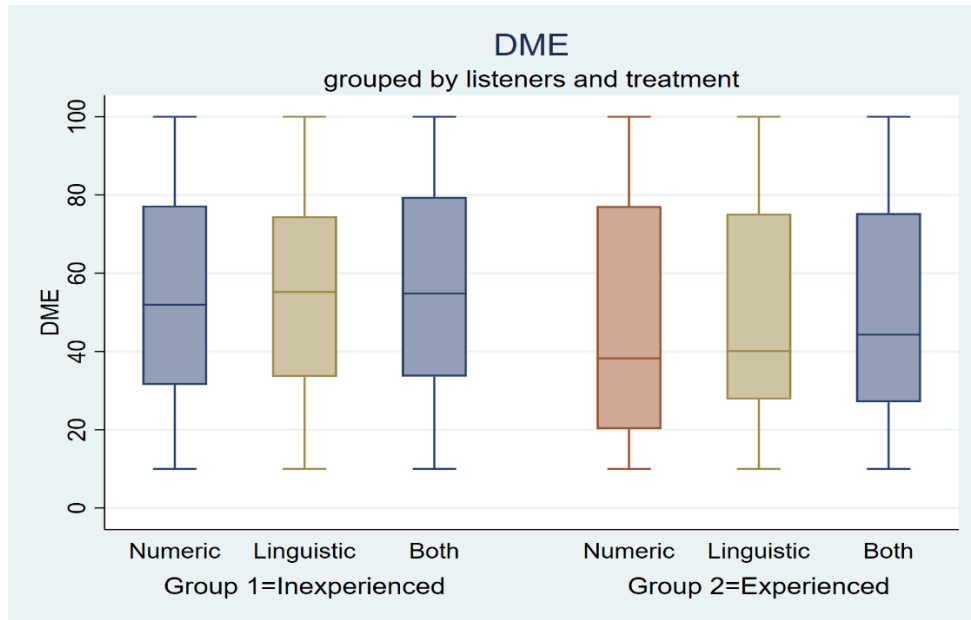


Figure 3-5. DME for both groups and all treatments.

Figure 3-5 facilitates the explanation of a few prominent facts. To elaborate, we colored the boxes for Linguistic treatments in Groups 1 and 2 light brown, and the box for the Numeric treatment in Group 2 a sienna color. We remark that for the Linguistic treatment, both groups have smaller dispersion than do Numeric and Both treatments. However, the mean is maintained among group treatments. Moreover, it is quite revealing that experts in Group 2 reported a larger SD (i.e., there was more error in each of the treatments as compared to Group 1). Table 3-7 indicates that experienced listeners' assertion of the DME carries even more error when performed only with a numeric scoring mechanism.

To maintain consistency with the established hypothesis regarding differences in variability of the estimates, we conducted Variance Ratio tests (see Tables 3-9 and 3-10). These tests allowed us to determine the equality of the SDs across treatments in a way similar to how t-tests determine differences in SDs between two samples (in this case, treatments). These treatments identify clear differences beyond the $P = 0.05$ significance level, specifically for tests against the

benchmark of the Linguistic treatment in Group 1. In other words, these tests substantiate the need for further analysis with control variables (i.e., a multivariate regression) to make improvements in DME estimation using linguistic assessments rather than the more commonly-accepted numeric assessment methods.

Table 3-9. Variance Ratio Test of Equality of Variance Inexperienced (Group 1)

Hypothesis	F-value	P-value
$SD(Numeric)=SD(Linguistic)$	1.287***	0.000
$SD(Numeric)=SD(Both)$	1.095	0.138
$SD(Linguistic)=SD(Both)$	0.851**	0.013

The variance is statistically higher in Group 1 (Numeric vs. Linguistic)

Table 3-10. Variance Ratio Test of Equality of Variance Experienced (Group 2)

Hypothesis	F-Value	P-Value
$SD(Numeric)=SD(Linguistic)$	1.120	0.060
$SD(Numeric)=SD(Both)$	1.173**	0.019
$SD(Linguistic)=SD(Both)$	1.047	0.478

Econometric Regression Analysis

Given that the treatment effects in the formalized tests prove that linguistic estimation methods produce the same mean DME estimates but with lower SD (i.e., less error), the next step is to conduct a regression analysis. The regression strategy is to make use of a modeling approach that accounts for repeated-measures data (i.e., across time). We use an appropriate cross-sectional estimation approach that corrects for autocorrelation in the presence of potential heteroskedasticity [99]. This is a relevant concern from an econometric standpoint because our data provide across-time variation (i.e., 100 audio recordings that provide repeated measures data). The model is linear

in the parameters, and beta results can be interpreted in a manner consistent with ordinary least squares (OLS) estimation, although the standard error estimates use the Newey and West correction [99]. The lag structure used was determined by visual analysis of the Partial Autocorrelation Function and the Autocorrelation Function plots of the respective dependent variables.

Summary statistics tables for each dependent variable and group are provided in Tables 3-11 and 3-12. As such, the observation in any given round covering either Group 1 or Group 2 is the SD of one treatment (e.g., Linguistic-only) minus the SD of the comparison treatment. The first model (Model 1) in both tables estimates the DME SD differences between Numeric and Linguistic treatments, and the second model (Model 2) in both tables estimates the DME SD differences between Both (Numeric/Linguistic) and Numeric treatments. Lastly, the third model (Model 3) in both tables estimates the DME SD differences between Both and Linguistic treatments.

Table 3-11. Summary of the Three Predictors Used in the Models for Inexperienced Listeners (Group 1)

Variable	Obs.	Mean	Std. Dev.	Min	Max
SD Diff: Numeric-Linguistic	100	2.448	5.256	-11.841	13.224
SD Diff: Numeric- Both	100	-.7424	5.119	-18.456	12.657
SD Diff: Linguistic- Both	100	-3.191	4.859	-17.480	7.911

Table 3-12. Summary of the Three predictors Used in the Models for Experienced Listeners (Group 2)

Variable	Obs.	Mean	Std. Dev.	Min	Max
SD Diff: Numeric-Linguistic	100	2.448	5.255	-11.841	13.224
SD Diff: Numeric- Both	100	1.105	7.243	-18.229	18.765
SD Diff: Linguistic- Both	100	-1.343	7.095	-20.457	13.652

Regression estimates are provided in Table 3-13 for three separate models corresponding to inexperienced listeners (Group 1). The same models are present in Table 3-14, but for experienced listeners (Group 2). Each regression contained identical right-hand variables, or the dummy variables representing each sound subtype. The non-contrast category was excluded as the reference category. The dependent variable (DV) in each model is the difference in DME SDs (SD) between any two of the three treatments (three models provide all three comparisons).

For regression Models 1 and 2 that include a difference comparison against the Linguistic-only treatment, the beta coefficients on the dummy variables provide insights into those sound categories (i.e., subtypes) that are most (or least) effectively assessed with linguistic scoring methods. In Model 1, because the DV equals the SD difference of Numeric minus Linguistic methods, a positive and statistically significant coefficient indicates that the variable (subtype) increases in variability, or SD, with the Numeric treatment relative to the Linguistic treatment. In other words, a positive sign means that the word category subtype has a relatively lower SD with Linguistic treatments than it does with Numeric treatments. The opposite is the case for negative coefficients that are statistically significant. For example, Table 3-14 shows that for high-low vowels, Linguistic treatments have higher variance (see Model 1 with -2.539 and Model 3 with 2.92, both with statistical significance).

Additionally, variance is higher for Both treatments, but is not higher for either Numeric or Linguistic treatments individually for fricative affricate under Group 2. Both Group 1 and Group 2 report very little difference between Both and Numeric-only treatments.

Statistically significant and positive coefficients are identified for Stop-affricate Consonants and Final Cluster-Final Singleton. Therefore, experimental results indicate that these word categories, or subtypes, are most effectively estimated using Linguistic scoring methods.

Statistically significant and negative coefficients are identified for only High-Low Vowels. For this subtype, higher variability is observed when using Linguistic assessment methods. Other positive and negative coefficients exist in the model, but they do not rise to commonly-accepted levels of statistical significance. Therefore, we do not interpret their differences as statistically different from zero, meaning that there may not be error improvement by using one assessment method over another for these subtype word categories.

The results from Model 3 provide interesting results, but not consistently in the expected direction. In Model 3, there are no statistically significant and positive coefficients. This indicates that the Linguistic assessment methods are not an improvement over assessment methods that combine Linguistic and Numeric, for any given subtype. However, both treatments represent an improvement over the Linguistic-only treatment in the presence of Fricative-Affricate Consonants, Alveolar-Palatal Consonants, High-Low Vowels, and Initial Cluster-Initial Singleton subtypes. These four subtypes each have negative and statistically significant dummies. For these word subtypes, the experimental results support the assertion that the most appropriate assessment methodology for measuring speech intelligibility is to provide listeners with both Linguistic and Numeric scoring assessment tools.

The regression estimates confirm the summary statistics in some compelling ways. The lowest difference in SD (i.e., the two treatments that are most similar in terms of DME value variability), an inconsistency equally higher when the listener is evaluating speech intelligibility using a Numeric assessment. However, variability is consistently lower (i.e., more precise and accurate) for Groups 1 and 2 in treatments where listeners can make use of Linguistic assessment methods.

There are some features of Linguistic terminology that are improving accuracy (i.e.,

reducing variability) in DME assessments. This lends support to Klir and Wierman [10] and Mendel [19], who suggest that uncertainty (i.e., that carried by *IntelliTurk*'s sliders while capturing a listener's assessment of the DME when presented with the Linguistic terms only treatment), "may often reduce complexity and, at the same time, increase the credibility of the model" [19, p. 256]. From this perspective, uncertainty plays a key role when modeling systems because it could be traded for gain in the other essential characteristics of the models [10, p. 4].

The similarity (or lack of difference) in the variability between Numeric and Both treatments also drives the lower model fitness measure (as provided by the F-statistic). Model 2, which provides this difference, is the only regression model that lacks a statistically-significant model fit. However, we include it for completeness. We regard Model 2 as less instructive because it is comparing variability across two assessment methodologies that each include the Numeric assessment of speech recordings, and are, therefore, unlikely to yield any real differences. The absence of this difference is, nonetheless, an important confirmation of our overarching hypothesis, because real improvement is exemplified in the Linguistic treatment.

Table 3-13. Regression Analysis Results Inexperienced Listeners (Group 1)

VARIABLES	(1)	(2)	(3)
	Numeric Minus Linguistic	Numeric Minus Both	Linguistic Minus Both
Stop_fricative	-1.058 (-0.480)	-0.117 (-0.0447)	0.941 (0.909)
Stop_affricate	4.610*** (3.403)	2.886*** (1.990)	-1.724 (-1.043)
Final_cluster_final_singleton	2.400*** (2.174)	2.434*** (2.598)	0.0341 (0.0395)
Fricative_affricate	-1.397 (-1.178)	1.611 (1.117)	3.008*** (3.115)
Alveolar_palatal_consonants	-1.990* (-1.451)	1.866 (1.140)	3.856** (1.754)
Front_back_vowels	1.566 (1.067)	1.962* (1.334)	0.396 (0.448)
High_low_vowels	-2.539***	1.150	3.690***

	(-2.042)	(0.671)	(2.671)
Initial_cluster_Initial_singleton	-1.468	2.339*	3.808***
	(-0.888)	(1.602)	(4.510)
Constant	2.611***	-2.428***	-5.039***
	(2.764)	(-2.880)	(-6.975)
Observations	100	100	100

Notes: Linear regression estimates with Newey-West t-statistics in parentheses. The dependent variable is the Difference (Diff) in treatment standard deviations in DMEs by round (100 rounds/audio recordings in total). *** p < 0.05, ** p < 0.10, * p < 0.20

Table 3-14. Regression Analysis Results Experienced Listeners (Group 2)

VARIABLES	(1)	(2)	(3)
	Numeric Minus Linguistic	Numeric Minus Both	Linguistic Minus Both
Stop_fricative	0.704 (2.516)	3.6241 (2.994)	3.624 (2.623)
Stop_affricate	4.690*** (1.752)	-1.644 (1.299)	-1.644 (1.655)
Final_cluster_final_singleton	1.643 (1.669)	0.758 (0.998)	0.758 (1.156)
Fricative_affricate	0.366 (1.237)	-1.717 (1.094)	-1.717 (1.270)
Alveolar_palatal_consonants	0.568 (1.502)	0.435 (1.145)	0.435 (1.327)
Front_back_vowels	1.204 (1.441)	1.704 (1.067)	1.704 (1.113)
High_low_vowels	-0.836 (1.664)	0.883 (1.625)	0.883 (1.861)
Initial_cluster_Initial_singleton	-1.042 (-0.888)	-0.539 (0.353)	-0.539 (0.936)
Constant	1.230 (1.193)	-1.214 (0.714)	-1.214 (0.776)
Observations	100	100	100

Notes: Linear regression estimates with Newey-West t-statistics in parentheses. The dependent variable is the Difference (Diff) in treatment standard deviations in DMEs by round (100 rounds/audio recordings in total). *** p < 0.05, ** p < 0.10, * p < 0.20

Moreover, as suggested by Mendel (2017), words mean different things to different people. However, when a system is designed that considers the appropriate number of alternatives, those other essential modeled characteristics of the system (i.e., the various word subtypes shown above) can benefit by leveraging the uncertainty inherent in the linguistic terms. For ease of interpretation,

the predictive margins of these statistically robust coefficients are shown in Tables 3-13 and 3-14. Plots of the predicted values from the regression estimation for each statistically-robust sound subtypes for Models 1 and 3 are presented below in Figs. 3-6, 3-7, 3-8, and 3-9, respectively. The x-axis values of each figure (i.e., in the “1” category on the right) provide the predicted difference in SD in the DME for stop-affricate sounds between the Numeric- and Linguistic-only treatments. The x-axis values on the left (i.e., the “0” category) represent the models’ predicted SD difference when all other word categories are held constant at their mean SDs. It can be interpreted simply as the mean, or average, SD that could be expected when any other given subtype is presented to an inexperienced listener. Regarding Model 1 Groups 1 and 2 (Figs. 3-6 and 3-8, respectively) and Model 3 Groups 1 and 2 (Figs. 3-7 and 3-9, respectively) margin plots, it is clear that Numeric-only assessment methods outperform Linguistic-only assessment methods for High-low Vowel sounds. Additionally, for stop-affricate and final cluster-final singleton sounds, it is clear that Linguistic-only assessments outperform Numeric-only assessments.

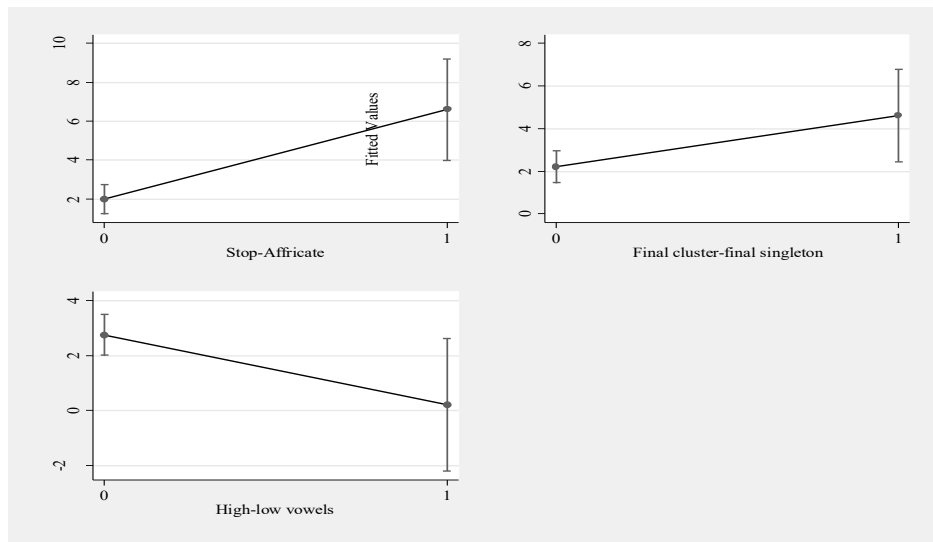


Figure 3-6. Margins plots (with 95% CIs) of subtypes with statistically-robust coefficients from Regression Model #1, Group 1.

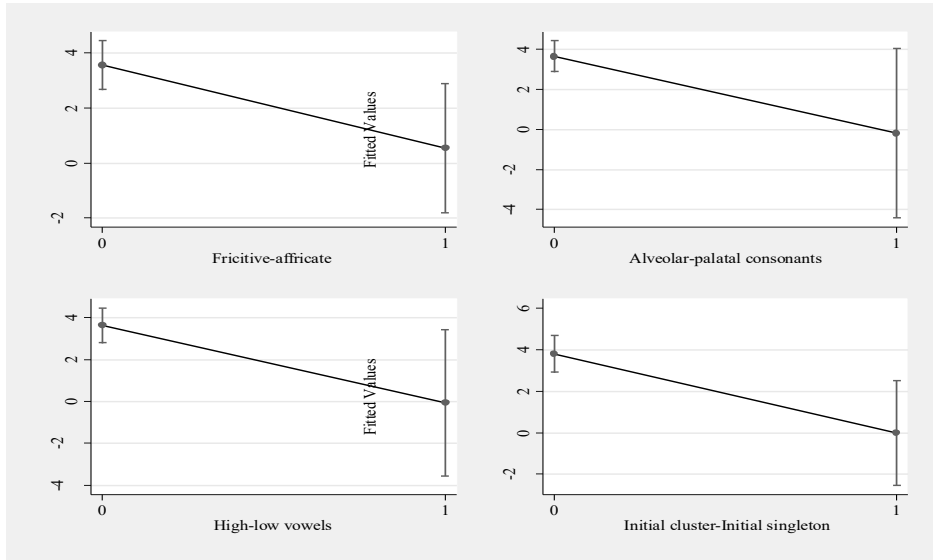


Figure 3-7. Margins plots (with 95% CIs) of subtypes with statistically-robust coefficients from Regression Model #3, Group 1.

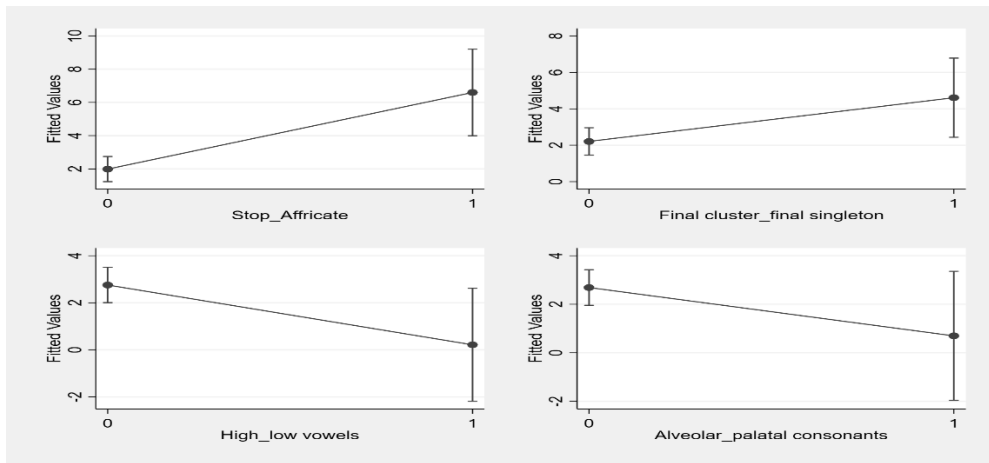


Figure 3-8. Margins plots (with 95% CIs) of subtypes with statistically-robust coefficients from Regression Model #1, Group 2.

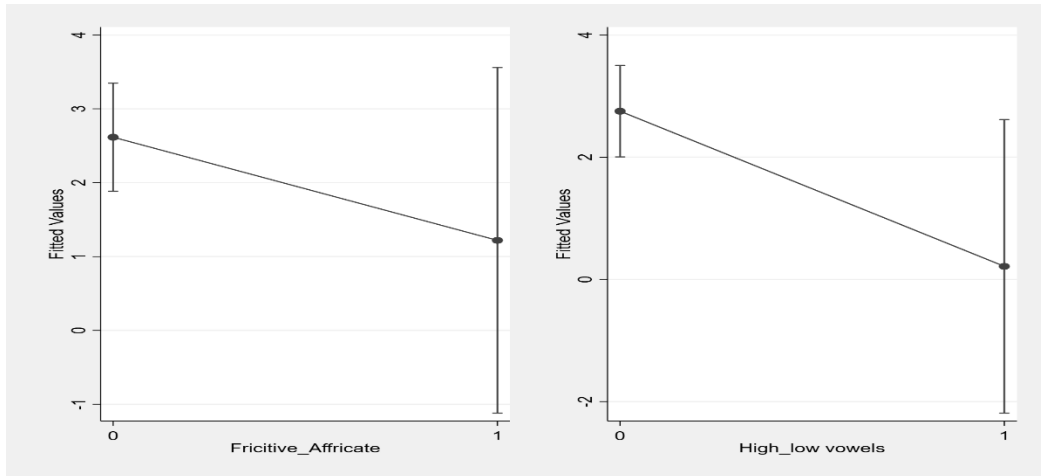


Figure 3-9. Margins plots (with 95% CIs) of subtypes with statistically-robust coefficients from Regression Model #3, Group 2.

Finally, we conducted an informative intra-model comparison. The inter-consistency suggests further validation of our findings across models we observed. We did not observe a “sign” change for any coefficient between any two models concerning the direction of comparison for a given assessment method. This means that, in all cases, the recommendations that would be inferred by clinicians regarding the experimental results (i.e., which word categories are most effectively assessed using which assessment methodology) are consistent regardless of the treatment group utilized. This intra-model comparison, which is essentially an intra-treatment comparison, suggests the following main takeaways. First and foremost, the main hypothesis is confirmed that linguistic assessments generally reduce error and improve the accuracy of intelligibility assessments. Second, some sound subtypes are more likely to be accurately assessed using one methodology over another. For stop-affricate and final cluster-final singleton subtypes, linguistic methods are likely to be more effective at reducing estimation error. For fricative-affricate, alveolar-palatal consonants, and initial cluster-initial singleton subtypes, providing the listener with both linguistic and numeric tools is likely to be more effective. Finally, for high-low

vowels, either numeric-only or numeric combined with linguistic assessment methods are likely to outperform linguistic-only assessment methods.

IntelliTurk Machine Learning Model

Machine learning is employed for this research, and Figure 3-10 shows the logical flow of the method used to incorporate deep machine learning into this study.

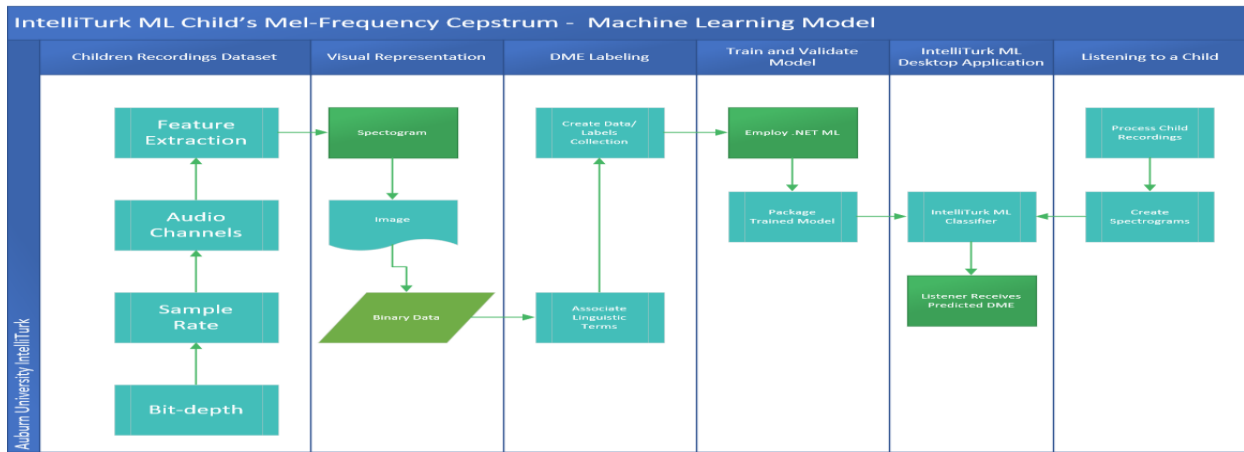


Figure 3-10. The Mel-Frequency Cepstrum Coefficients (MFCC) logical flow for how recordings are converted into spectrograms (images).

Mel-Frequency Cepstrum Coefficients (MFCC)

MFCC is a feature extraction method based on how humans perceive sounds' frequencies through their senses [100]. This algorithm has been used by researchers to study crying infants with hypothyroidism [101]. MFCC makes it possible to convert a sound recording's binary data into an image. This process facilitates the use of deep learning algorithms that perform high accuracy classifications [4] [7] [53], making it possible to classify DMEs as a result of listener assertion of child intelligibility.

Researchers at the MathWorks Lab, makers of MATLAB and Simulink, have highlighted that starting in 2016, deep learning continues to out-perform humans on image classification [102].

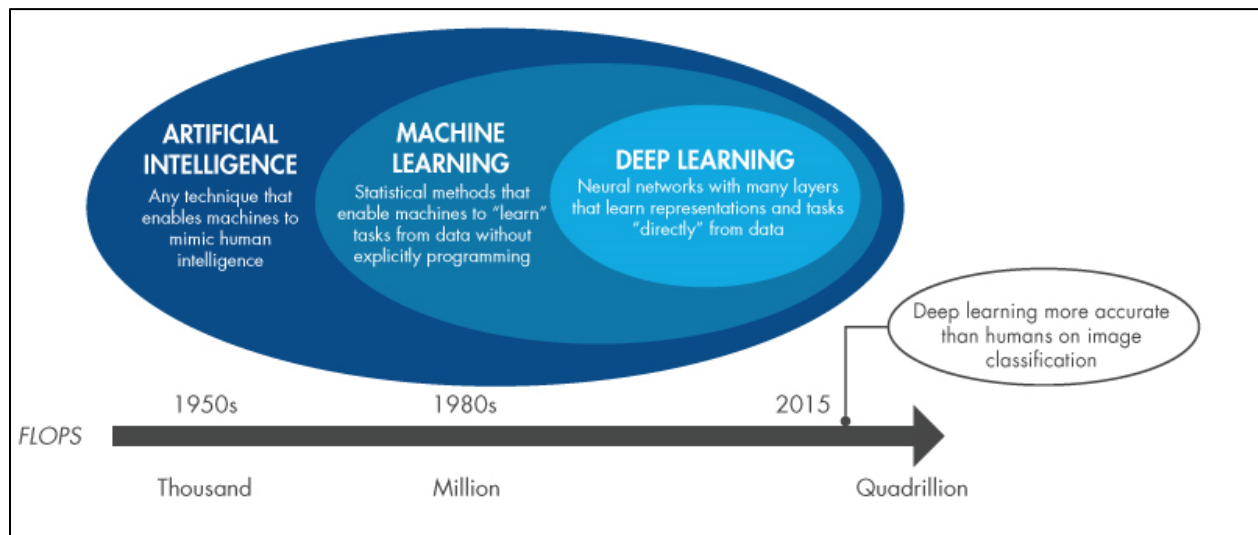


Figure 3-11. MathWorks depicts deep learning as a subset of machine learning, which is a subset of artificial intelligence [103].

Moreover, the MathWorks Lab’s researchers highlight the tremendous growth in available data over the years, starting with thousands of data points in the 1950s, growing to a million in the 1980s, and remarkably reaching a quadrillion in 2015 [102]. As such, Supervised Learning, another subset of machine learning where collected data is labeled and used to train and validate models, is employed to solve classifications (predictions) problems. Nonetheless, deep learning is a subset of machine learning [103], and deep learning’s core is “Neural Networks with many layers that learn representations and tasks directly from data” [53]. Moreover, researchers have designed deep learning algorithms (e.g., convolutional neural networks) for extracting images’ features with state-of-the-art classification accuracy [104].

Singh and Ghosh stated, "deep learning are very powerful tool to develop various mechanisms for health care management, especially in disease diagnosis for stroke care." [1]. Moreover, Patterson and Gibson suggested using deep learning when there is complex pattern matching in images (i.e., spectrograms) [2] and such statement aligns with [105] and others like [1], [5] and, [101].

In this work, deep learning is leveraged as a tool that improves the modeling of DME to assist trained clinicians in assessing intelligibility in children. Deep learning serves as an effective estimation approach for conceptualizing the problem of estimating a child's speech intelligibly beyond the listener's subjective intuition.

For modeling the DME, we choose Microsoft .NET ML TensorFlow implementation, an opensource and cross-platform machine learning framework [106] [107]. This library implements the original Google Brain architecture, as shown in Figure 3-5. Moreover, Table 3-3 shows the breakdown of the steps we followed to successfully train a deep neural network (DNN) capable of predicting the DME.

We converted the voice recordings of children into spectrograms (images; see Table 3-15). Step one indicated in Table 3-16 involves incorporating a specialized library (cs-mel-spectrogram) [108] that processes recordings (Figure 3-10, "Listening to Child Recordings") and produces the MFCC coefficients.

Dataset

DME values captured from the experiments on experts are split into three subsets, one for model training, one for model validation, and another for model testing. After keeping a very small percentage of the data to test the model, the remainder of it will be used for training and validation.

We used 80% of the dataset for training the model and 20% to validate it. Regarding avoiding invalid results, in particular from the training dataset, even when experiments are repeated and the results are validated, Abu-Mostafa *et al.* [52, p. 172] remark “if the data is sampled in a biased way, learning will produce a similarly biased outcome.” Additionally, “if the dataset has affected any step in the learning process, its ability to assess the outcome has been compromised” [52, p. 172].

Data Preprocessing and Feature Extraction

To ensure consistency across the whole dataset, we preprocessed the following audio properties: accuracy. Moreover, a visual representation of each recorded word audio file assisted us in identifying these features for classification (prediction) using the same techniques employed to classify images with high accuracy. We used the MFCC technique that is similar to spectrograms and used for visualizing the spectrum of frequencies of a sound and how they vary during a short time. MFCC functions similarly to how the human auditory system processes sound. Figure 3-10 depicts the process of converting .wav files into spectrograms. Spectrograms use a linear spaced frequency scale (so, each frequency bin is spaced an equal number of Hertz apart), and MFCC uses a quasi-logarithmic spaced frequency scale.

The Deep Learning Model (DNN)

We built and trained a DNN with the above datasets, and used it to make DME predictions. DNN’s typically make good classifiers and perform particularly well with image classification tasks due to their feature extraction and classification accuracy. DNNs are effective at finding patterns within the MFCCs much like they are effective at finding patterns within images.

We trained the model starting with a low number of epochs and a low batch size. If we observed from the output that the model was converging, we increased both epochs and batch size, because training a DNN can take a significant amount of time. The next step is to review the accuracy of the model using both the training and test datasets.

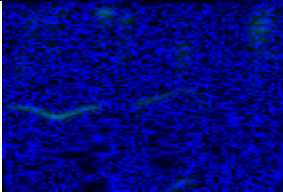
Lab Validation

One of our goals is to have the *IntelliTurk*-DME Predictor Tool used by experts for real lab work validation. As such, we aim for a confidence level of at least 80%. In essence, DNN model performance was estimated in terms of its accuracy at predicting the DME of experienced listeners using new recordings of children’s words.

Each of the expert’s evaluations, either in agreement or disagreement with the DNN model prediction, was recorded. Model performance was determined using the ratio of matchings (i.e., experts agreed with predicted values) to the total number of predictions.

Table 3-15 shows a subset of the spectrograms that result from converting the sound recordings of children. Note that the density of the blue area is less evenly distributed as the level of difficulty increases; for example, a “Very Easy” DME classification of the word “television” is shown by an evenly distributed spectrogram, as compared to a “Very Difficult” DME classification for the word "stairs.”

Table 3-15. Spectrograms Images from Child Word Recordings

Spectrogram	DME Classification	Child Recorded Word
	Very Easy	Television

	Easy	Tomato
	Medium	Geese
	Difficult	Mouse
	Very Difficult	Stairs

The MFCCs are the extracted features (from the children’s voice recordings) used for the DNN model. Some of these features include audio-channels, sample rate, and bit-depth, as shown in Figure 3-3 “Children Recording Dataset.”

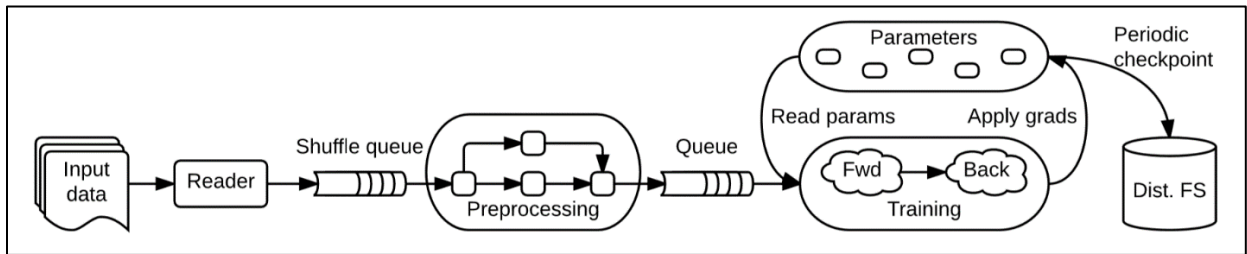


Figure 3-12. Google Brain Schematic TensorFlow Dataflow graph for a training pipeline [109]

Table 3-16. Steps to Get a Deep Learning Model for the Prediction of DME Based on Google Brain’s TensorFlow

Step	Description
1	Convert children recordings into spectrograms (images)

2	Load the initial full image-set (spectrograms) into an IDataView, the input and output of Query Operators (Transforms). This is the fundamental data pipeline type, and shuffle so it will be better balanced.
3	Load Images with in-memory type within the IDataView and Transform Labels to Keys (Categorical)
4	Split the data 80:20 into train (80%) and test (20%) sets, train and evaluate.
5	Define the model's training pipeline using DNN (TensorFlow library for .Net) default values
6	Train/create the ML model - Measuring training time
7	Collect quality metrics (accuracy, etc.)
8	Save the model to assets/outputs

Table 3-17. Portion of Resources Representing Child Recordings Used to Score DME

ID	Gender	Age	Age in Months	Age in Years	Ethnicity	Native Language	Race	Speech Status	Total Records
2AU201-01DM4_9SSD	Male	4;9	57	4.75	NA	English	African-American	SSD	55
2AU201-02NF6_8	Female	6;8	80	6.67	NA	English	African-American	NSSD	56
2AU201-07DM3_0SSD	Male	3;0	36	3	Not Hispanic or Latino	English	White	SSD	54
2AU201-08DF2_0SSD	Female	2;0	24	2	Not Hispanic or Latino	English	White	SSD	54
2AU201-09DF4_0SSD	Female	4;0	48	4	Not Hispanic or Latino	English	White	SSD	11
2AU201-10NF4_5	Female	4;5	53	4.42	Not Hispanic or Latino	English	White	NSSD	204
2AU201-11NF3_8	Female	3;8	44	3.67	Not Hispanic or Latino	English	White	NSSD	214
2AU201-12NM3_10	Male	3;10	46	3.83	Not Hispanic or Latino	English	White/African American	NSSD	178
2AU201-14NF4_10	Female	4;10	58	4.83	Not Hispanic or Latino	English	White	NSSD	212

Note: Number of records indicate number of recordings associated with a child

Table 3-18. DME Scores Given by Trained Clinicians and Collected by the IntelliTurk.ML Windows-Based Application

Seed Break Down ID	SLP	DME Term
21	SLP Listener1	Very Easy

166	SLP Listener1	Medium
8	SLP Listener1	Difficult
55	SLP Listener1	Very Easy
24	SLP Listener1	Very Difficult
691	SLP Listener2	Medium
649	SLP Listener2	Medium
744	SLP Listener2	Very Easy
575	SLP Listener2	Difficult
667	SLP Listener2	Easy
614	SLP Listener2	Very Difficult
779	SLP Listener2	Very Easy
740	SLP Listener2	Very Easy
610	SLP Listener2	Easy
540	SLP Listener2	Difficult

Furthermore, as shown in Table 3-15, the features extracted from the spectrogram’s images were fed into a DNN model following the flow shown in Figure 3-13. Table 3-18 shows an excerpt from the DME scores given by trained clinicians and collected by a custom Windows-based application specifically designed and coded for this task.

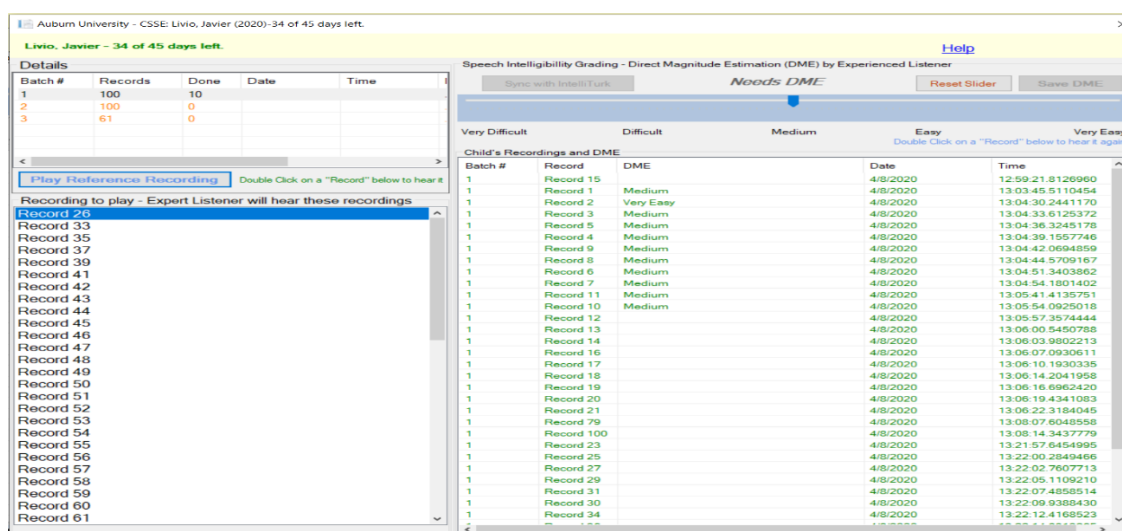


Figure 3-13. A Windows-based application used by trained clinicians for listening to child word recordings and scoring their intelligibility.

Regarding the steps listed in Table 3-16, a DNN core principle is guiding the training of intermediate levels of representation using unsupervised learning, which can be performed locally at each level. These levels use a learning algorithm that greedily trains one layer at a time, exploiting an unsupervised learning algorithm for each layer (i.e., a Restricted Boltzmann Machine, RBM).

Deep Learning, based on a layer-wise-greedy-learning algorithm was proposed by [6], leverages unsupervised learning as a step needed to extract features (pretraining) before the neural network trains layer-by-layer, thus, “by extracting features from the inputs, the data dimension is reduced and a compact representation is hence obtained” [53] [105]. Once the features are extracted and the sample data points are labeled, they are passed down to the next layer where they are processed as refined fuel (i.e., de-errored data).

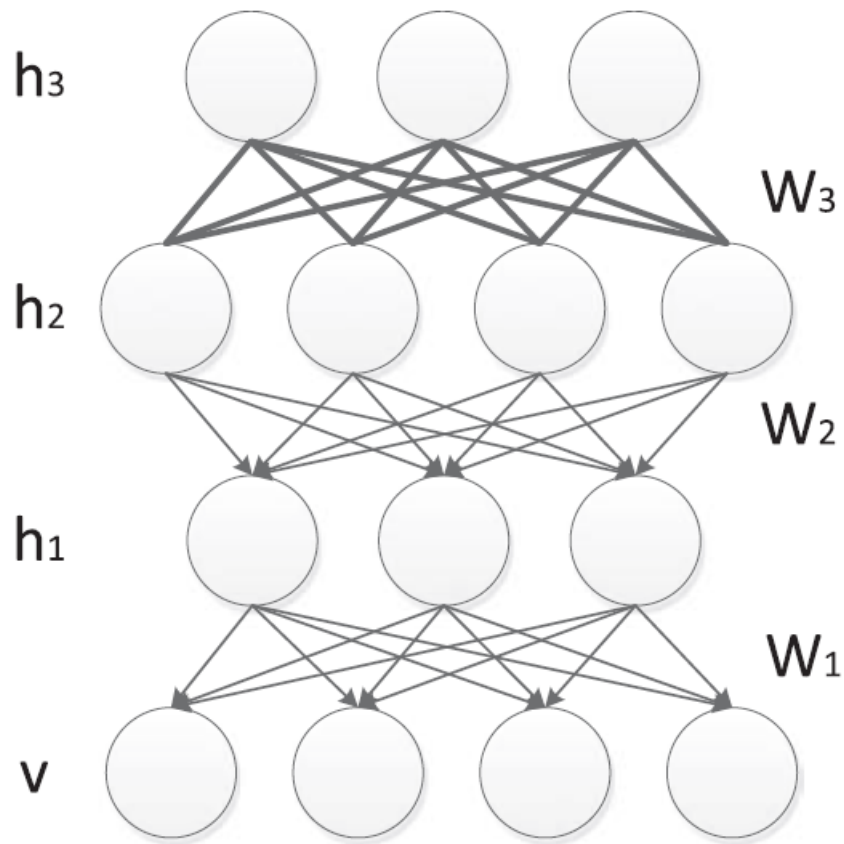


Figure 3-14. Core structure of a Deep Neural Network (DNN), as originally proposed by Hinton [105]

In a DNN, the network's initial weights are learned from the structure of the input data (i.e., feature learning), moving away from the previously used stochastic gradient descent that was known for its tendency to lead to over- or under-fitting [105].

Moreover, the DNN has shown high accuracy when modeling tasks like the human visual system and image classification [105], thus making it an excellent candidate for classifying the DME based on spectrogram images, as shown in Table 3-15.

When features are extracted by the top RBM, they are propagated back to the lower layers. As the DNN extracts a spectrogram's features, it sends them to lower layers. Therefore, the lower

layers train first, followed by the layer above in a top-down flow, as shown in Figure 3-14 [105] [6].

Preparing and Testing the DNN

The DNN model was trained, validated, and deployed as a compressed file into a Windows-based application (see Appendix 3B), it shows both the training and validation steps as rendered in Table 3-16. The model was deployed and a prediction was made for the spectrogram image corresponding to a child recorded word of “bed” (see Appendices 3D). This spectrogram was produced from a recording heard by a trained clinician who scored the child’s intelligibly with a DME level equal to “Easy.” This file was not included during the training nor during the validation of the DNN model.

The DNN prediction accuracy equals the “proportion of correct predictions with a test data set. It is the ratio of the number of correct predictions to the total number of input samples” [110]. For this recording of “bed,” prediction accuracy was approximately 41% (i.e., the maximum of the five probability scores (see Appendix 3D): 0.06123, 0.4134, 0.2203, 0.2382, and 0.0666).

Furthermore, once the model was deployed, another prediction was conducted using a Windows-based application designed and coded to test the DNN trained model upon its deployment, thus simulating a production environment, as shown in Figure 3-15. Here, the DNN was 52% accurate for a word classified as “Medium” by a trained clinician for the child recording “2AU201-35NF5_7-PC65_Medium.wav.” This recording was not part of the training or validation processes of the DNN model.

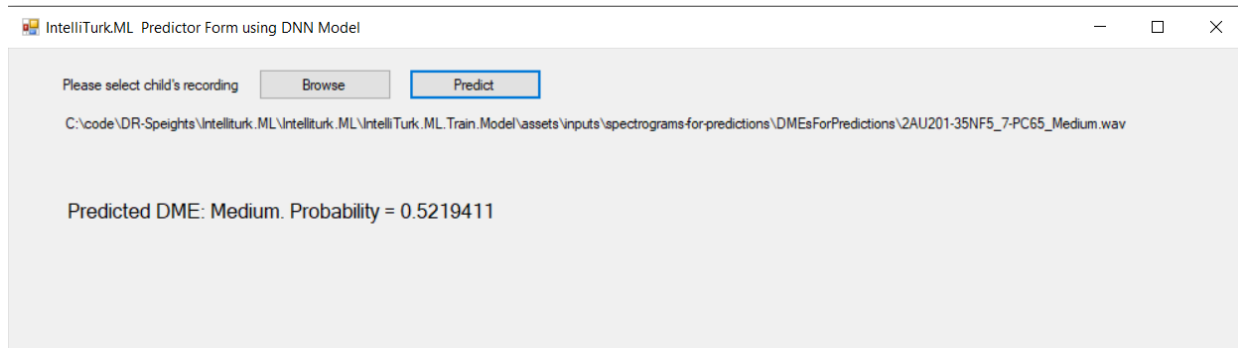


Figure 3-15. Testing the DNN-deployed model with classification of a spectrogram image employing a Windows-based application.

This section raises some questions that we address in the following sections. For example, a deep learning model such as the *IntelliTurk* DNN requires considerable amounts of data. Additionally, due to the complexity of converting children's voice recordings into spectrograms, employing the MFCC algorithm (i.e., it converts speech signal as a sequence of spectral vectors) to represent the speech signal as a time versus frequency charts, the yielded images showed dark regions at higher amplitudes (see Table 3-15 for examples). The darker the spectrogram the harder for the DNN to learn it.

The Efficacy of the *IntelliTurk* Model Supported through DNN

We built the *IntelliTurk* POC, conducted several experiments (as seen in Table 3-15), and gathered and analyzed data, as shown in Tables 3-13 and 3-14. However, is this study worth continuing, and are we moving in the right direction? We have gathered evidence that the DNN is learning from the expert participants, but the use of more experts will strengthen our method in future iterations. Additionally, we have a limitation regarding the current environment (i.e., windows-based application) and not being able to gather data from as many experts as planned.

However, we used a sufficient number of experts to substantiate that our model is being trained successfully.

The amount data used to train the *IntelliTurk* DNN indicates that the middle range holds a very strong variability, dispersion based on the complexity of child intelligibility metrics (DMEs) as reflected in the spectrograms' features (Table 3-15). Moreover, to analyze the model's capacity to generalize (i.e., inferring DME scores from recordings that were not used to train the model), the DME metrics (that were transformed into spectrograms) were grouped by linguistic terms according to the following: Group 1: Very-Difficult and Very-Easy; Group 2: Difficult and Easy; Group 3: Very-Difficult, Difficult, and Medium; Group 4: Medium, Easy, and Very-Easy; Group 5: Very-Difficult, Difficult, Easy, and Very-Easy; Group 6: Very-Difficult, Difficult, Medium, Easy, and Very-Easy).

The models describe that Group 1 (which excludes middle DME metrics) has the highest accuracy. Figure 3-16 shows the DNN with over 80% accuracy during training and validation. Moreover, Group 2 (which also excludes middle DME metrics) renders the DNN with close to 50% accuracy, which is less than Group 1 accuracy by around 30%. Furthermore, as the middle metrics are included in the DNN training and validation (as observed in Figs. 3-18 and 3-19), it is noticeable that model accuracy fluctuates drastically (Figure 3-19). Representing Group 5, Figure 3-20 shows the DNN with very low accuracy. Additionally, we observed that the DNN including all available data (i.e., Group 6) experienced a consistent decrease in both training and validation loss and had a model accuracy of around 21% (Figure 3-21). However, we found that the model is successfully generalizing based on the experts' DME determinations.

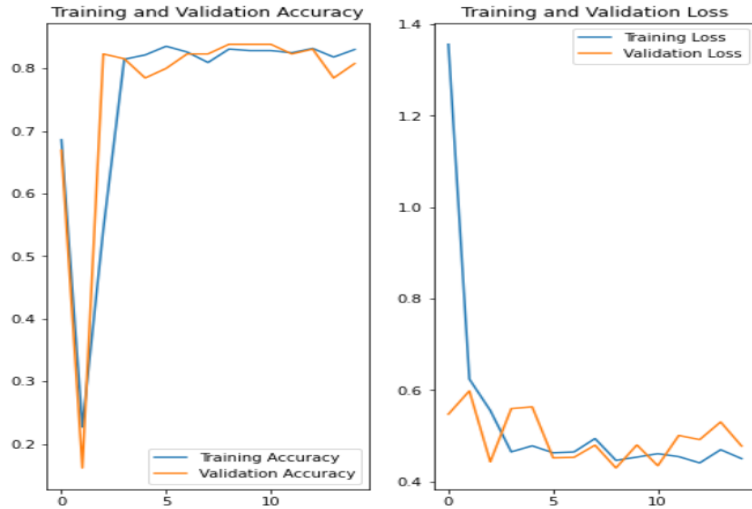


Figure 3-16. DNN model including Very-Difficult and Very-Easy metrics (i.e., Group 1).

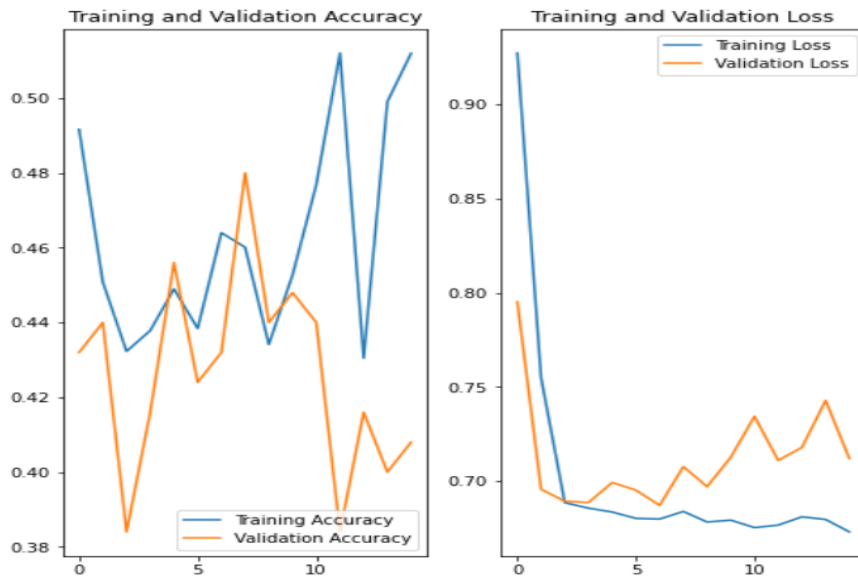


Figure 3-17. DNN model including Difficult and Easy metrics (i.e., Group 2).

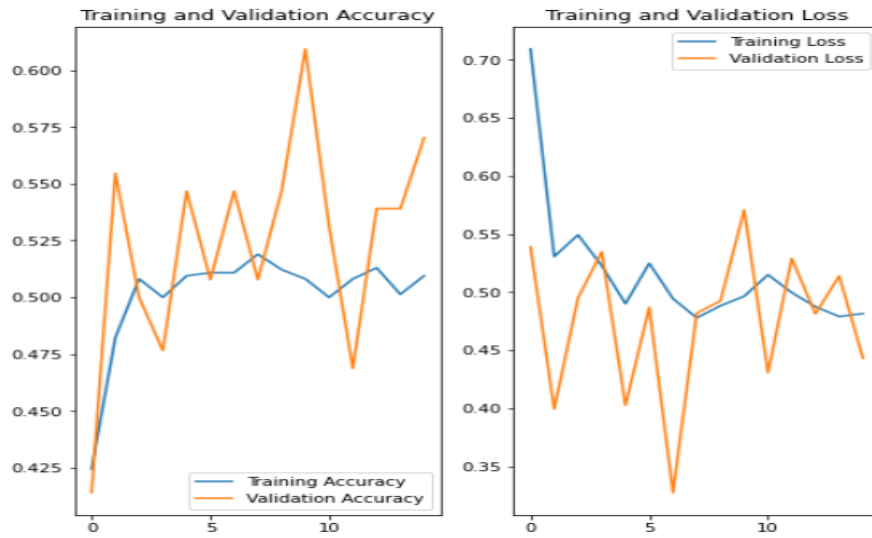


Figure 3-18. DNN model including Very-Difficult, Difficult, and Medium metrics (i.e., Group 3).

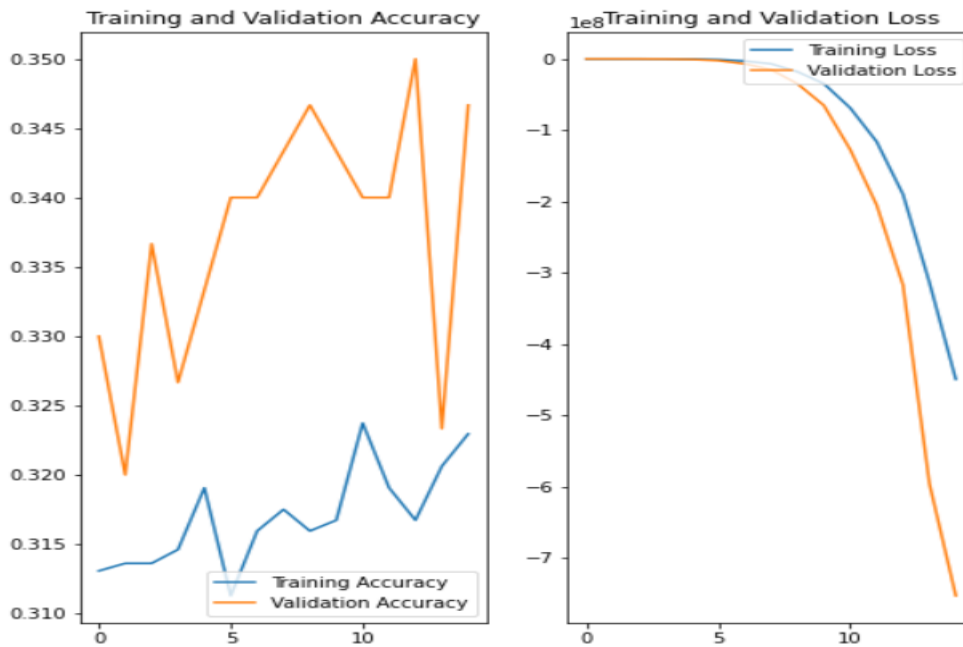


Figure 3-19. DNN model including Medium, Easy, and Very-Easy metrics (i.e., Group 4).

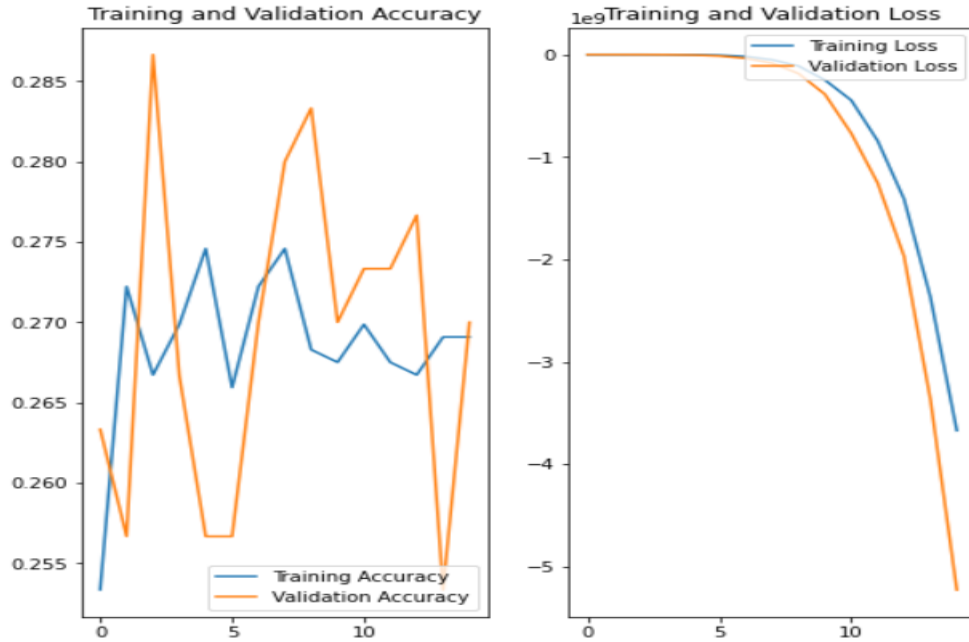


Figure 3-20. DNN model including Very-Difficult, Difficult, Easy, and Very-Easy metrics (i.e., Group 5).

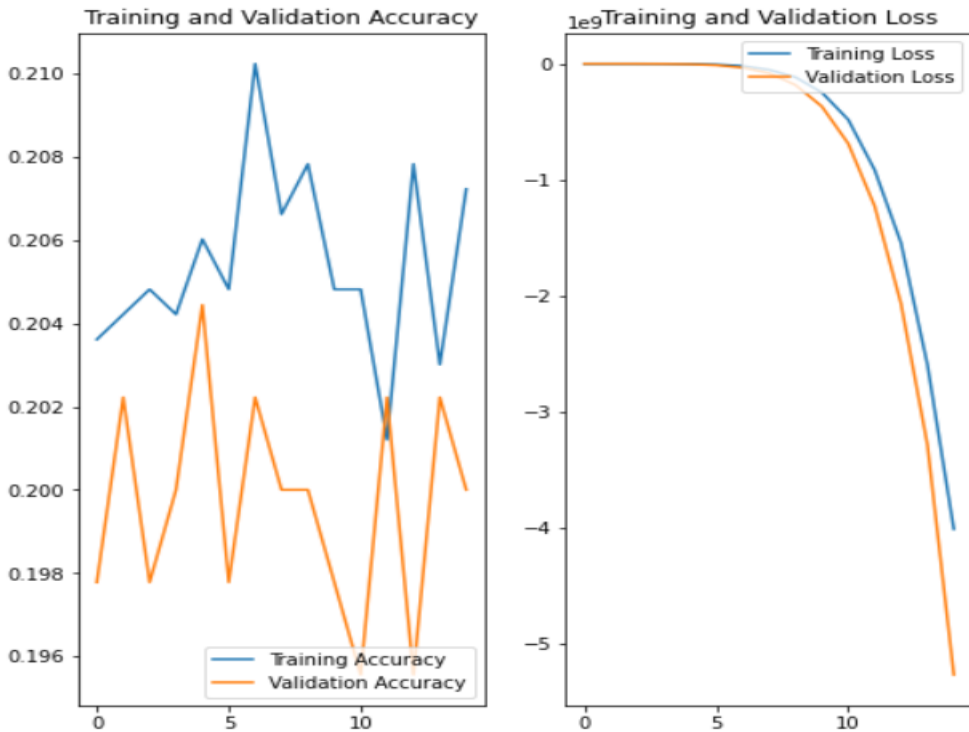


Figure 3-21. DNN model including Very-Difficult, Difficult, Medium, Easy, and Very-Easy metrics (i.e., Group 6).

The use of the MFCC algorithm to produce spectrograms from children's word recordings provides opportunities to investigate specific acoustic features needed to enhance DNN accuracy. Moreover, DME metrics fall into the category of subjective measurements of perceptual phenomenon (i.e., listening to a child's voice), and our results indicate that DME metrics in the middle ranges were more challenging for the DNN to learn. This has been previously observed for researchers in the field of perceptual learning. For example, a group of researchers have proven that metrics in the middle ranges are very noisy (i.e., hard to detect). Additionally, when presented with a greater number of response options, humans do not tend to choose the responses at either end of the scale, but rather tend to select metrics from the center [111] [112].

Conclusions and Future Work

Results indicate the presence of a treatment effect. Additionally, there is consistency in that the DME value is statistically the same whether a subject observes a numeric or linguistic value, according to Figs. 3-5 and 3-6. We conclude that the numeric procedure does not inflate estimates and that the true estimate is captured by linguistic terms, which do not suppress or deflate the DME. However, Tables 3-7 and 3-8 show that Numeric methods produce a larger SD, while Linguistic methods produce less dispersion. This could be evidence that linguistic terms carry less uncertainty, as they convey human stimuli upon perceiving them. Moreover, when subjects are presented with both numeric and linguistic methods, DME values are inflated and more dispersed (Table 3-12). Figure 3-8 shows an increasing trend captured by regression model 3 (Group 1), while Figures. 3-9 and 3-6 show a slightly decreasing line captured by regression models 3 (Group 2) and 1 (Group 1), respectively.

Additionally, a variance analysis indicates that 100 questions do not fatigue subjects. Fatigue is generally consistent for the first and last 25 questions, and at the same time, the last ten questions as well. The detractor is that the variance is statistically different across treatments (which is good) but this is not the same for earlier versus later questions. Hence, something is fatiguing subjects—we think this may be consistent with one of our main hypotheses. For this reason, we think this potentially evidences that it takes the subject less work/effort to report linguistic versus numeric values and in later rounds, the variance in the linguistic treatment is less. So, the stress-reduction benefits (potentially) of linguistic terminology are borne out more in later rounds but this is not observed in early rounds.

Panel models show robust treatment differences with some notable caveats. The Both treatment (i.e., Numeric and Linguistic) is statistically significant in most models, but the others (e.g., Numeric and Linguistic) are not as significant. Nonetheless, since capturing numeric or linguistic values resulted in the same DME value and that, through the progress of the experiments, subjects seemed to show fatigue only when using numeric DME, we reinforce the idea of replacing numeric with linguistic methods, which hold less dispersion.

In several studies, phonetic contrast types were found to have different contributions to a listener's ability to understand the speech signal. Prior studies in child speech disorders have contributed a listener's impressions of decreased intelligibility to a deviation related to specific phonetic contrast error types [113]. The findings from our current study are consistent with those reported in prior literature in that listeners experience more difficulty with the consonantal contrast categories of stop-affricate, fricative-affricate, alveolar-palatal, and high-low vowels, as well syllable structure categories of final cluster-final singleton and initial cluster-initial singleton subtypes. Our study adds to this body of literature by postulating that the mechanism with which

this perceptual construct is measured for child speech intelligibility can affect not only the rating's variability, but may contribute to the identification of different phonetic categories ascribed to the perceived decrease in intelligibility. Hodson and Paden [93] identified syllable reduction as those that affect clusters that have a greater impact on intelligibility. The results from an unpublished pilot study support that the aforementioned phonetic contrast types are more challenging for listeners [114]. The increased difficulty in measuring intelligibility for these categories may benefit from the use of both Numerical and Linguistic modalities of scoring.

Our study indicates that linguistic methods generally reduce error and improve the accuracy of intelligibility assessments. Additionally, we found that some sound subtypes are likely to be more accurately assessed using one methodology over another. Moreover, when needing to further explain some observed categorical effects, like those of stop-affricate and final cluster-final singleton subtypes, linguistic methods are likely to be more effective at reducing estimation error. For fricative-affricate, alveolar-palatal consonants, and initial cluster-initial singleton subtypes, providing the listener with a combination of linguistic and numeric assessment methods is likely to improve accuracy. Moreover, for high-low vowels, either Numeric-only or Both assessment methods are likely to outperform Linguistic-only methods. Further investigation is warranted, however, as some scholars have highlighted that fuzzy systems employ linguistic terms [38] [80] [81] that are primarily used in deductive reasoning (or shallow reasoning) where the specific is inferred from the general [29, pp. 8-9]. Inductive reasoning (i.e., inferring the general from the particular) and deep reasoning (i.e., capturing those processes of mother nature that produce phenomenon) do not perform well when modeling complex systems, such as listening to a child to determine its intelligibility. We are currently developing a machine learning model that employs deep learning, a subset of artificial intelligence proven to perform well when modeling functions

not otherwise efficiently representable. When approached statistically, we expect poor generalization due to a deep architecture that is insufficient for representing these functions given the amount of data available [4] [19], a limitation that we expect to improve as more data become available. Additionally, automatic speech recognition has been challenged by child speech, and further investigation is warranted in identifying acoustic features beyond the MFCC for the robust classification of child speech intelligibility.

Furthermore, the *IntelliTurk* web-based application could evolve into a micro-service, cloud-based application. This will make *IntelliTurk* a more robust platform that supports both desktops and mobile devices.

References

- [1] S. Ghosh and A. Singh, "The scope of Artificial Intelligence in mankind: A detailed review," *Journal of Physics: Conference Series*, vol. 1531, no. 012045, 2020.
- [2] J. Patterson and A. Gibson, *Deep Learning*, O'Reilly Media, Inc., 2017.
- [3] G. Montibeller and D. Von Winterfeldt, "Cognitive and Motivational Biases in Decision and Risk Analysis," *Risk Analysis*, vol. 35, no. 7, pp. 1230-1252, 2015.
- [4] H. Larochelle, D. Erhan and P. Vincent, "Deep Learning using Robust Interdependent Codes," pp. 312-319, 2008.
- [5] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Sardinia, 2010.
- [6] G. E. Hinton, S. Osindero and Y.-W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation*, vol. 18, no. 7, pp. 1527-1554, 2006.
- [7] B. Póczos, "Advanced Introduction to Machine Learning, CMU-10715 Deep Learning," 2017.
- [8] J. Y. Halpern, *Reasoning About Uncertainty*, Second ed., Cambridge, Massachusetts: The MIT Press, 2017.
- [9] B. C. Ezzell, "Infrastructure Vulnerability Assessment Model (I-VAM)," *Risk Analysis*, vol. 27, no. 3, pp. 571-583, 2007.
- [10] G. Klir and M. Wierman, *Uncertainty-Based Information. Elements of Generalized Information Theory*, Heidelberg, Berlin: Springer-Verlag, 1999.
- [11] K. Mu, J. Zhi, R. Lu and W. Liu, "Measuring Inconsistency in Requirements Specifications," in *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty: Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, Aalborg, Denmark, 2005.
- [12] G. J. Klir, *Uncertainty and Information. Foundations of Generalized Information Theory*, Hoboken: John Wiley & Sons, Inc., 2006.
- [13] I. Sommerville, *Software Engineering*, Ninth Edition, London: Pearson, 2009.
- [14] D. Dubois and H. Prade, "Possibility Theory and its Applications: Where Do we Stand?," IRIT-CNRS, Université Paul Sabatier, Toulouse, 2011.
- [15] L. A. Zadeh, "The Concept of Linguistic Variable and its Application to Approximate Reasoning," *Elsevier*, vol. 8, no. Information Sciences, pp. 199-249, 1975.
- [16] T. J. Ross, *Fuzzy Logic With Engineering Applications*. Third Edition., Chippenham, Wilshire: CPI Antony Rowe, 2010.
- [17] J. J. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic Theory and Applications*, New Jersey: Prentice Hall PTR, 1995.
- [18] E. Rutherford, "The Chemical Nature of the Alpha Particles from Radioactive Substances," The Nobel Prize, 11 12 1908. [Online]. Available: <https://www.nobelprize.org/prizes/chemistry/1908/rutherford/lecture/>. [Accessed 29 6 2020].
- [19] J. M. Mendel, *Uncertain Rule-Based Fuzzy Systems Introduction and New Directions*, Second Edition ed., Los Angeles, California: Springer, 2017.
- [20] D. J. Koehler and N. Harvey, *Blackwell Handbook of Judgment and Decision Making*, Malden, MA: Blackwell Publishing Ltd, 2004.

- [21] A. K. Shah and D. M. Oppenheimer, "Heuristics Made Easy: An Effort-Reduction Framework," *American Psychological Association*, vol. 134, no. 2, pp. 207-222, 2008.
- [22] Ohio Public Utilities Commission, "Direct Testimony of Noah Dormady on Behalf of the Office of the Ohio Consumers' Counsel electronically filed by Ms. Deb J. Bingham on behalf of Healey, Christopher Mr.," 2 1 2019. [Online]. Available: <https://dis.puc.state.oh.us/TiffToPDF/A1001001A19A02B24524H00207.pdf>. [Accessed 18 7 2020].
- [23] D. W. Hubbard, *How to Measure Anything Finding the Value of "Intangibles" in Business*, Third ed., New Jersey, New Jersey: Wiley & Sons, Inc., 2014.
- [24] Startkey, "Livio AI Reimagining what a hearing aid can do," [Online]. Available: <https://www.starkey.com/hearing-aids/livio-artificial-intelligence-hearing-aids>. [Accessed 20 7 2020].
- [25] Noom, "Make the world a healthier place," Noom, [Online]. Available: <https://www.noom.com/>. [Accessed 20 7 2020].
- [26] R. Schmelzer, "How to build a machine learning model in 7 steps," *Cognilytica*, 1 7 2020. [Online]. Available: <https://searchenterpriseai.techtarget.com/feature/How-to-build-a-machine-learning-model-in-7-steps>. [Accessed 17 7 2020].
- [27] D. Jung, "Recent Machine Learning Applications in Space," *IEEE Potentials*, pp. 34-38, 6 07 2020.
- [28] D. Dubois and H. Prade, *Possibility Theory An Approach to Computerized Processing of Uncertainty*, USA, Ed., New York, New York: Plenum Press, 1988.
- [29] T. J. Ross, *Fuzzy Logic With Engineering Applications*, Third Edition., Chippenham, Great Britain: John Wiley & Sons, Ltd., 2010.
- [30] L. A. Zadeh, "Fuzzy set as a basis for a theory of possibility," *Elsevier Science*, vol. 100, no. 1, pp. 3-28, 1978.
- [31] A. A. Alola, M. Tunay and V. Alola, "Analysis of Possibility Theory for Reasoning under Uncertainty," *Internation Journal of Statatistics and Probability*, vol. 2, no. 2, pp. 12-23, 2013.
- [32] J. A. Livio, W. Flores, R. Hodhod and D. Humphress, "Smart Fuzzy Cupper: Employing approximate reasoning to derive coffee bean quality scoring from individual attributes," in *IEEE World Congress On Computational Intelligence, WCCI 2018*, Rio de Janeiro, Brazil, 2018.
- [33] J. A. Livio and R. Hodhod, "AI Cupper: A Fuzzy Expert System for Sensorial Evaluation of Coffee Bean Attributes to Derive Quality Scoring," *IEEE Transactions on Fuzzy Systems*, vol. Early Access, no. DOI: 10.1109/TFUZZ.2018.2832611, pp. 1-10, 2018.
- [34] D. Dubois, W. Liu, J. Ma and H. Prade, "The basic principle of uncertain information fusion. An organised review of merging rules in different representation frameworks," *ELSEVIER*, vol. 32, pp. 12-39, 2016.
- [35] A. Engelbrecht, *Computational Intelligence an Introduction*, West Sussex: John Wiley & Sons Ltd, 2007.
- [36] E. Cox, *The Fuzzy Systems Handbook*, San Diego: AP Professional , 1999.
- [37] L. A. Zadeh, "Outline of a New Approach to the Analysis of Complex Systems and Decision Processes," *IEEE Trans. Systems, Man, and Cybernetics*, no. 1, pp. 28-44, 01 1973.
- [38] M. Negnevitsky, *Artificial Intelligence, A Guide to Intellegent Systems*, Third ed., Harlow: Addison Wesley, 2011.

- [39] J. Rivera, "Unlocking Coffee's Chemical Composition: Part 2," *Coffee Chemistry*, 06 5 2015. [Online]. Available: <https://www.coffeechemistry.com/library/coffee-science-publications/unlocking-coffee-s-chemical-composition-part-2>. [Accessed 1 12 2017].
- [40] W. C. Flores and G. M. Pineda, "A Type-2 Fuzzy Logic System Approach to Train Honduran Coffee Cuppers," in *Computational Intelligence (LA-CCI), 2016 IEEE Latin American Conference on, Cartagena, Colombia, 2-4 Nov. 2016*. doi: 10.1109/LA-CCI.2016.7885710..
- [41] E. Cox, *The Fuzzy Systems Handbook*, Second ed., Chappaqua, New York: AP PROFESSIONAL, 1999.
- [42] SCAA, "SCAA Protocols, Cupping Specialty Coffee," Santa Ana, 2013.
- [43] Cafe Culture International, "The Q&A of Q Grading," Cafe Culture International, 26 4 2013. [Online]. Available: <http://www.cafeculture.com/industrynews/the-qa-of-q-grading>. [Accessed 12 11 2016].
- [44] J. Ma, "A correspondence between belief function combination and knowledge base merging," *Intentional Journal of Approximate Reasoning*, vol. 104, pp. 1-8, 2018.
- [45] A. P. Dempster, "Upper and Lower Probabilities Induced by a Multivalued Mapping," in *Studies in Fuzziness and Soft Computing*, R. R. Yager and L. Liu, Eds., Springer, 2008, pp. 57-72.
- [46] P. Smets and R. Kennes, "The transferable belief model," *Artificial Intelligence*, vol. 66, no. 1065, pp. 191-234.
- [47] J. Lin, "Integration of weighted knowledge bases," *Artificial Intelligence*, vol. 83, no. 2, pp. 363-378, June 1996.
- [48] C. D. Tupper, "Interpreting Models," *Data Architecture From Zen to Reality*, pp. 281-305, 2011.
- [49] R. Kruse, J. Heinsohn and R. Schwecke, *Uncertainty and Vagueness in Knowledge Based Systems Numerical Methods*, Berlin Heidelberg: Springer-Verlag, 1991.
- [50] A. M. Borghi, *Grounding cognition: The role of perception and action in memory, language, and thinking*, D. Pecher and R. A. Zwaan, Eds., Cambridge University Press, 2005.
- [51] L. McClendon and N. Meghanathan, "Using machine learning algorithms to analyze crime data," *Machine Learning and Applications: An International Journal (MLAIJ)*, vol. 2, no. 1, pp. 1-12, 2015.
- [52] Y. S. Abu-Mostafa, M. Magdon-Ismail and H.-T. Lin, *Learning From Data a Short Course*, Online: AML Book, 2012.
- [53] Y. Bengio, "Learning Deep Architectures for AI," *Association for Computing Machinery - Foundations and Trends® in Machine Learning*, vol. 2, no. 1, pp. 1-130, 2009.
- [54] N. Bankson, J. Bernthal and P. Flipsen, "Speech Sound Assessment Procedures," in *Articulation and phonological disorders: Speech sound disorders in children*, 2013, pp. 180-211.
- [55] R. Perren, R. Geiger, S. Schenker and F. Escher, "Recent Developments in Coffee Roasting Technology," *Research Gate*, pp. 451-459, 2015.
- [56] W. Flores and G. M. Pineda, "A Type-2 Fuzzy Logic System Approach to Train Honduran Coffee Cuppers," in *Computational Intelligence (LA-CCI)*, Cartagena - Colombia, 2016.
- [57] J. A. Livio and R. Hodhod, "AI Cupper: A Fuzzy Expert System for Sensorial Evaluation of Coffee Bean Attributes to Derive Quality Scoring," *IEEE Transactions on Fuzzy Systems*, vol. Early Access, pp. 1-10, 2018.
- [58] B. Marr, "A Short History of Machine Learning -- Every Manager Should Read," 19 2 2016. [Online]. Available: <https://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read/#5a29f04315e7>. [Accessed 12 3 2018].

- [59] P. S. Prudvi and E. Sharifahmadian, "Applying machine learning techniques to find important attributes for heart failure severity assessment," *International Journal of Computer Science, Engineering and Applications (IJCSA)*, vol. 7, no. 5, 2017.
- [60] P. Vithu and J. A. Moses, "Machine vision system for food grain quality evaluation: A review," *ELSEVIER: Trends in Food Science & Technology*, vol. 56, pp. 13-20, 2016.
- [61] Coffee Research Institute, "Coffee Roasting," Coffee Research Institute, 2006. [Online]. Available: <http://www.coffeeresearch.org/coffee/roasting.htm>. [Accessed 1 12 2017].
- [62] W. Godoy Fontes , I. Nunes da Silva, A. Goedel and H. R. Cunha Palácios, "Fuzzy Logic Applied at Industrial Roasters in the Temperature Control," in *11th IFAC Workshop on Intelligent Manufacturing Systems. The International Federation of Automatic Control*, São Paulo, Brazil, 2013.
- [63] T. J. Ross, *Fuzzy Logic With Engineering Applications*. Third Edition., Chippenham, Wilshire: CPI Antony Rowe, 2010.
- [64] K. Nogueira, W. R. Schwartz and J. A. Dos Santos, "Coffee Crop Recognition Using Multi-scale Convolutional Neural Networks," Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, 2014.
- [65] T. Bevis, *Java Design Pattern Essentials - Second Edition*, Leigh-on-Sea: Ability First Limited, 2012.
- [66] Wikipedia, "Principal Component Analysis," Wikipedia The Free Encyclopedia, 24 7 2020. [Online]. Available: https://en.wikipedia.org/wiki/Principal_component_analysis. [Accessed 29 7 2020].
- [67] L. Breiman, *Random Forests*, Berkeley: Statistics Department University of California Berkeley, CA 94720, 2001.
- [68] T. A. Santhanam and M. B. Padmavathi, "Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis," *Procedia Computer Science*, vol. 47, pp. 76-83, 2015.
- [69] M. K. Sohrabi and A. Tajik, "Multi-objective feature selection for warfarin dose prediction," *Computational Biology and Chemistry*, vol. 69, no. PMID:28689108, pp. 126-133, 2017.
- [70] J. H. Connolly, "Intelligibility: a linguistic view," *Intentional Journal of Language & Communication Disorders*, vol. 21, no. 3, pp. 371-376, 1986.
- [71] K. C. Hustad, A. Sakash, P. E. Natzke, A. T. Broman and P. J. Rathouz, "Longitudinal Growth in Single Word Intelligibility Among Children With Cerebral Palsy From 24 to 96 Months of Age: Predicting Later Outcomes From Early Speech Production," *Journal of Medicine and Speech Pathology*, vol. 20, no. 4, pp. 1599-1613, 2019.
- [72] K. Hustad, A. Oakes and K. Allison, "Variability and diagnostic accuracy of speech intelligibility scores in children," *Journal of Speech, Language, and Hearing Research*, vol. 58, no. 6, pp. 1695-1707, 2015.
- [73] S. M. Skahan, M. Watson and G. L. Lof, "Speech-Language Pathologists' Assessment Practices for Children With Suspected Speech Sound Disorders: Results of a National Survey," *American Journal of Speech-Language Pathology*, vol. 16, pp. 246-259, 09 2007.
- [74] S. Stevens and E. Galanter, "Ratio scales and category scales for a dozen perceptual continua. Journal of experimental psychology," *Journal of experimental psychology*, vol. 54, no. 6, pp. 377-411, 1957.
- [75] N. Schiavetti, *Scaling procedures for the measurement of speech intelligibility*, R. D. Kent, Ed., Madison, Wisconsin: John Benjamins Publishing Company, 1992, pp. 11-34.

- [76] N. Schiavetti, D. Metz and R. Sitler, "Construct validity of direct magnitude estimation and interval scaling of speech intelligibility: Evidence from a study of the hearing impaired.," *Journal of Speech and Hearing Research*, vol. 24, no. 3, pp. 441-445, 1981.
- [77] G. Weismer and J. S. Laures, "Direct Magnitude Estimates of Speech Intelligibility in Dysarthria: Effects of a Chosen Standard," *Journal of Speech, Language, and Hearing Research*, vol. 45, pp. 421-433, 06 2002.
- [78] S. s. Stevens, "The Direct Estimation of Sensory Magnitudes: Loudness," *The American Journal of Psychology*, vol. 69, no. 1, pp. 1887-2019, 1956.
- [79] J. Liss, S. Spritzer, J. N. Caviness and C. Adler, "The effects of familiarization on intelligibility and lexical segmentation in hypokinetic and ataxic dysarthria," *The Journal of the Acoustical Society of America*, vol. 112, no. 6, pp. 3022-3030, 2002.
- [80] J. J. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic Theory and Applications*, New Jersey: Prentice Hall PTR, 1995.
- [81] E. Cox, *The Fuzzy Systems Handbook*, Second ed., Chappaqua, New York: AP PROFESSIONAL, 1999.
- [82] K. M. Allison, "Measuring Speech Intelligibility in Children With Motor Speech Disorders," *Perspectives of the ASHA Special Interest Groups*, pp. 1-12, 2020.
- [83] S. H. Han, M. Song and J. Kwahk, "A systematic method for analyzing magnitude estimation data," *International Journal of Industrial Ergonomics*, pp. 513-524, 1997.
- [84] R. T. Verrillo, "Stability of line-length estimates using the method of absolute magnitude estimation," *Perception and Psychophysics*, pp. 261-265, 1983.
- [85] J. Feldmand and J. C. Baird, "Magnitude estimation of multidimensional stimuli," *Perception & Psychophysics*, pp. 418-422, 1971.
- [86] F. H. Petzschner and S. Glasauer, "A Bayesian perspective on magnitude estimation," *Trends in Cognitive Sciences*, 2015.
- [87] AMAZON, "Access a global, on-demand, 24x7 workforce," 9 8 2019. [Online]. Available: <https://www.mturk.com/>.
- [88] M. Becker and J. Levine, "Experigen – an online experiment platform," 2013. [Online]. Available: <http://becker.phonologist.org/experigen/>. [Accessed 2 11 2019].
- [89] "A framework for creating linguistic experiments," 18 9 2014. [Online]. Available: <https://github.com/tlozoot/experigen>. [Accessed 2 11 2019].
- [90] S. Kawahara, "Psycholinguistic Methodology in Phonological Research," *Oxford Bibliographies*, pp. 1-20, 1 2016.
- [91] W. Williams, D. Zhou, G. Stewart and P. Knott, "The practicality of using a smart phone 'App' as an SLM and personal noise exposure meter (SoundLog)," *2nd Australasian Acoustical Societies Conference, ACOUSTICS*, vol. 2016, no. 1, pp. 547-553, 2016.
- [92] W. Secord and J. A. S. Donohue, "CAAP: Clinical Assessment of Articulation and Phonology," *Super Duper Publications*, 2002.
- [93] B. W. Hodson and E. P. Paden, "Targeting intelligible speech: A phonological approach to remediation," in *College-Hill Press*, Austin, TX, 1983.
- [94] B. Dodd, H. Zhu, A. Crosbie and A. Ozanne, *Diagnostic evaluation of articulation and phonology*, Psychology Corporation., 2002.

- [95] D. Ingram, "The measurement of whole-word productions," *Journal of Child Language*, vol. 29, no. 4, pp. 713-733, 2002.
- [96] M. Ross and J. Lerman, "Word Intelligibility by Picture Identification," *ERIC*, p. 46, 1971.
- [97] K. M. Cienkowski, M. Ross and J. Lerman, The Word Intelligibility by Picture Identification (WIPI) Test Revisited, Storrs, Connecticut: University of Connecticut,, 1971, pp. 39-43.
- [98] M. C. Coppens-Hofman, H. Terband, A. F. Snik and B. A. Maassen, "Speech Characteristics and Intelligibility in Adults with Mild and Moderate Intellectual Disabilities," *Folia Phoniatrica Et Logopaedica*, vol. 68, no. 4, pp. 175-182, 25 1 2017.
- [99] W. K. Newey and K. D. West, "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, vol. 55, no. 3, pp. 703-708, 1987.
- [100] A. Zabidi, W. Mansor, L. Y. Khuan, I. M. Yassin and R. Sahak, "Investigation of Mel Frequency Cepstrum Coefficients Parameters for Classification of Infant Cries with Hypothyroidism using MLP Classifier," in *Neural Networks (IJCNN), International Joint Conference on*, Barcelona, Spain, 2010.
- [101] C. Kim On, P. M. Pandiyan, S. Yaacob and A. Saudi, "Mel-frequency cepstral coefficient analysis in speech recognition," in *2006 International Conference on Computing & Informatics*, Kuala Lumpur, 2006.
- [102] MathWorks, "Terminology," The MathWorks, Inc., 2020. [Online]. Available: <https://explore.mathworks.com/machine-learning-vs-deep-learning/chapter-1-129M-100NU.html>. [Accessed 30 5 2020].
- [103] MathWorks, Inc., "Deep Learning or Machine Learning?," 2020. [Online]. Available: <https://explore.mathworks.com/machine-learning-vs-deep-learning/chapter-1-129M-100NU.html>. [Accessed 11 6 2020].
- [104] MathWorks, Inc., "Your Data," MathWorks, Inc., 2020. [Online]. Available: <https://explore.mathworks.com/machine-learning-vs-deep-learning/chapter-3-129M-99NU.html>. [Accessed 5 7 2020].
- [105] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing - Elsevier*, vol. 234, pp. 11-26, 2016.
- [106] Microsoft, "ML.NET An open source and cross-platform machine learning framework," Microsoft, 2020. [Online]. Available: <https://dotnet.microsoft.com/apps/machinelearning-ai/ml-dotnet>. [Accessed 12 6 2020].
- [107] Microsoft, "TensorflowCatalog Class," 2020. [Online]. Available: <https://docs.microsoft.com/en-us/dotnet/api/microsoft.ml.tensorflowcatalog?view=ml-dotnet>. [Accessed 12 6 2020].
- [108] C. Caishun, "cschen1205," 4 2018. [Online]. Available: <https://github.com/cschen1205/cs-mel-spectrogram>. [Accessed 12 6 2020].
- [109] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu and X. Zheng, "TensorFlow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, Atlanta, 2016.
- [110] Microsoft, "Evaluate your ML.NET model with metrics," 17 12 2019. [Online]. Available: <https://docs.microsoft.com/en-us/dotnet/machine-learning/resources/metrics>. [Accessed 14 6 2020].

- [111] Massachusetts Institute of Technology, "How expectation influences perception," 15 7 2019. [Online]. Available: <http://news.mit.edu/2019/how-expectation-influences-perception-0715>. [Accessed 23 7 2020].
- [112] C. Vinney, "Likert Scale: What Is It and How to Use It?," 3 6 2019. [Online]. Available: <https://www.thoughtco.com/likert-scale-4685788>. [Accessed 23 7 2020].
- [113] R. D. Kent, G. Weismer, J. F. Kent and J. C. Rosenbek, "Toward Phonetic Intelligibility Testing in Dysarthria," *Journal of Speech and Hearing Disorders*, vol. 54, pp. 482-499, 1989.
- [114] M. Speights Atkins, K. E. Willoughby, A. J. Weaver , M. J. Sandage, D. J. Bailey and J. A. Livio Zavarce, "A Typological Approach to the Examination of Word Recognition Errors in Crowdsourced Speech Intelligibility Determinations," *American Journal of Speech Language Pathology (submitted)*, 2020.

Appendix 2A. SCA Cupping Form



Specialty Coffee Association of America Coffee Cupping Form

Name: _____
Date: _____

Quality scale:			
6.00 - Good	7.00 - Very Good	8.00 - Excellent	9.00 - Outstanding
6.25	7.25	8.25	9.25
6.50	7.50	8.50	9.50
6.75	7.75	8.75	9.75

Sample #	Roast Level or sample	Score: <input type="text"/>	Score: <input type="text"/>	Score: <input type="text"/>	Score: <input type="text"/>	Score: <input type="text"/>	Score: <input type="text"/>	Score: <input type="text"/>	Score: <input type="text"/>	Total Score	
		Fragrance/Aroma	Flavor	Acidity	Body	Uniformity	Clean Cup	Overall			
		6 7 8 9 10	6 7 8 9 10	6 7 8 9 10	6 7 8 9 10	6 7 8 9 10	6 7 8 9 10	6 7 8 9 10	6 7 8 9 10		
		Dry Qualities Break	Aftertaste Score: <input type="text"/>	Intensity High Low	Level Heavy Thin	Balance Score: <input type="text"/>	Sweetness Score: <input type="text"/>	Defects (subtract) Taint=2 # cups Intensity Fault=4 <input type="text"/> X <input type="text"/> = <input type="text"/>			
Notes: _____										Final Score	

Sample #	Roast Level or sample	Score: <input type="text"/>	Score: <input type="text"/>	Score: <input type="text"/>	Score: <input type="text"/>	Score: <input type="text"/>	Score: <input type="text"/>	Score: <input type="text"/>	Score: <input type="text"/>	Total Score	
		Fragrance/Aroma	Flavor	Acidity	Body	Uniformity	Clean Cup	Overall			
		6 7 8 9 10	6 7 8 9 10	6 7 8 9 10	6 7 8 9 10	6 7 8 9 10	6 7 8 9 10	6 7 8 9 10	6 7 8 9 10		
		Dry Qualities Break	Aftertaste Score: <input type="text"/>	Intensity High Low	Level Heavy Thin	Balance Score: <input type="text"/>	Sweetness Score: <input type="text"/>	Defects (subtract) Taint=2 # cups Intensity Fault=4 <input type="text"/> X <input type="text"/> = <input type="text"/>			
Notes: _____										Final Score	

Sample #	Roast Level or sample	Score: <input type="text"/>	Score: <input type="text"/>	Score: <input type="text"/>	Score: <input type="text"/>	Score: <input type="text"/>	Score: <input type="text"/>	Score: <input type="text"/>	Score: <input type="text"/>	Total Score	
		Fragrance/Aroma	Flavor	Acidity	Body	Uniformity	Clean Cup	Overall			
		6 7 8 9 10	6 7 8 9 10	6 7 8 9 10	6 7 8 9 10	6 7 8 9 10	6 7 8 9 10	6 7 8 9 10	6 7 8 9 10		
		Dry Qualities Break	Aftertaste Score: <input type="text"/>	Intensity High Low	Level Heavy Thin	Balance Score: <input type="text"/>	Sweetness Score: <input type="text"/>	Defects (subtract) Taint=2 # cups Intensity Fault=4 <input type="text"/> X <input type="text"/> = <input type="text"/>			
Notes: _____										Final Score	

Appendix 2B: SCA Protocol, Roasting and Sample Preparation

Sample Preparation

Roasting:

- The sample should be roasted within 24 hours of cupping and allowed to rest for at least 8 hours.
- Roast profile should be a light to light-medium roast, measured via the M-Basic (Gourmet) Agtron scale of approximately 58 on whole bean and 63 on ground, +/- 1 point (55-60 on the standard scale or Agtron/SCAA Roast tile #55).
- The roast should be completed in no less than 8 minutes and no more than 12 minutes. Scorching or tipping should not be apparent.
- Sample should be immediately air-cooled (no water quenching).
- When they reach room temperature (app. 75° F; 23° C), completed samples should then be stored in airtight containers or non-permeable bags until cupping to minimize exposure to air and prevent contamination.
- Samples should be stored in a cool dark place, but not refrigerated or frozen.

To determine measurement:

- The optimum ratio is 8.25 grams per 150 ml of water, as this conforms to the mid-point of the optimum balance recipes for the Golden Cup.
- Determine the volume of water in the selected cupping glass and adjust weight of coffee to this ratio within +/- .25 grams.

Cupping Preparation:

- Sample should be ground immediately prior to cupping, no more than 15 minutes before infusion with water. If this is not possible, samples should be covered and infused not more than 30 minutes after grinding.
- Samples should be weighed out AS WHOLE BEANS to the predetermined ratio (see above for ratio) for the appropriate cup fluid volume.
- Grind particle size should be slightly coarser than typically used for paper filter drip brewing, with 70% to 75% of the particles passing through a U.S. Standard size 20 mesh sieve. At least 5 cups from each sample should be prepared to evaluate sample uniformity.
- Each cup of sample should be ground by running a cleansing quantity of the sample through the grinder, and then grinding each cup's batch individually into the cupping glasses, ensuring that the whole and consistent quantity of sample gets deposited into each cup. A lid should be placed on each cup immediately after grinding.

Pouring:

- Water used for cupping should be clean and odor free, but not distilled or softened. Ideal Total Dissolve Solids are 125-175 ppm, but should not be less than 100 ppm or more than 250 ppm.
- The water should be freshly drawn and brought to approximately 200° F (93°C) at the time it is poured onto the ground coffee.

Appendix 2C: SCA Protocol, Roasting and Sample Evaluation

Sample Evaluation

Sensory testing is done for three reasons:

- To determine the actual sensory differences between samples
- To describe the flavor of samples
- To determine preference of products

No one test can effectively address all of these, but they have common aspects. It is important for the evaluator to know the purpose of the test and how results will be used. *The purpose of this cupping protocol is the determination of the cupper's preference.* The quality of specific flavor attributes is analyzed, and then drawing on the cupper's previous experience, samples are rated on a numeric scale. The scores between samples can then be compared. Coffees that receive higher scores should be noticeably better than coffees that receive lower scores.

The Cupping Form provides a means of recording 11 important flavor attributes for coffee: Fragrance/Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Clean Cup, Sweetness, Defects, and Overall. The specific flavor attributes are positive scores of quality reflecting a judgment rating of the cupper; the defects are negative scores denoting unpleasant flavor sensations; the Overall score is based on the flavor experience of the individual cupper as a personal appraisal. These are rated on a 16-point scale representing levels of quality in quarter point increments between numeric values from 6 to 9. These levels are:

Quality Scale

6.00 - Good	7.00 - Very Good	8.00 - Excellent	9.00 - Outstanding
6.25	7.25	8.25	9.25
6.50	7.50	8.50	9.50
6.75	7.75	8.75	9.75

Theoretically the above scale ranges from a minimum value of 0 to a maximum value of 10 points. The lower end of the scale is below specialty grade.

Evaluation Procedure

Samples should first be visually inspected for roast color. This is marked on the sheet and may be used as a reference during the rating of specific flavor attributes. The sequence of rating each attribute is based on the flavor perception changes caused by decreasing temperature of the coffee as it cools:

Appendix 2D: SCA Protocol, Sample Evaluation Steps

Step #1 – Fragrance/Aroma

1. Within 15 minutes after samples have been ground, the dry fragrance of the samples should be evaluated by lifting the lid and sniffing the dry grounds.
2. After infusing with water, the crust is left unbroken for at least 3 minutes but not more than 5 minutes. Breaking of the crust is done by stirring 3 times, then allowing the foam to run down the back of the spoon while gently sniffing. The Fragrance/Aroma score is then marked on the basis of dry and wet evaluation.

Step #2 – Flavor, Aftertaste, Acidity, Body, and Balance

3. When the sample has cooled to 160° F (about 70° C), 8-10 minutes from infusion, evaluation of the liquor should begin. The liquor is aspirated into the mouth in such a way as to cover as much area as possible, especially the tongue and upper palate. Because the retro nasal vapors are at their maximum intensity at these elevated temperatures, Flavor and Aftertaste are rated at this point.
4. As the coffee continues to cool (160° F - 140° F; 70° C - 60° C), the Acidity, Body and Balance are rated next. Balance is the cupper's assessment of how well the Flavor, Aftertaste, Acidity, and Body fit together in a synergistic combination.
5. The cupper's preference for the different attributes is evaluated at several different temperatures (2 or 3 times) as the sample cools. To rate the sample on the 16-point scale, circle the appropriate tick-mark on the cupping form. If a change is made (if a sample gains or loses some of its perceived quality due to temperature changes), re-mark the horizontal scale and draw an arrow to indicate the direction of the score.

Step #3 – Sweetness, Uniformity, and Cleanliness

6. As the brew approaches room temperature (below 100° F; 37 ° C) Sweetness, Uniformity, and Clean Cup are evaluated. For these attributes, the cupper makes a judgment on each individual cup, awarding 2 points per cup per attribute (10 points maximum score).
7. Evaluation of the liquor should cease when the sample reaches 70° F (21° C) and the Overall score is determined by the cupper and given to the sample as "Cupper's Points" based on ALL of the combined attributes.

Step #4 – Scoring

After evaluating the samples, all the scores are added as describe in the "Scoring" section below and the Final Score is written in the lower right hand box.

Appendix 2E: Coffees' Origin, Roasting Data and Their Final Quality Grading Scores

1	SampleCode	COUNTRY	INITIALTEMP	MINTEMP	MINSECS	MAXTEMP	MAXSECS	FINALTEMP	SECONDS	SCORE
2	CO14	COLOMBIA	383	176	28	406	546	406	547	89
3	ET01	ETHIOPIA	367	167	29	396	616	394	622	86.79
4	CO09	COLOMBIA	399	178	32	401	592	401	604	86.46
5	HO01	HONDURAS	376	172	31	392	640	392	650	86.07
6	ES04	EL SALVADOR	381	181	33	410	675	288	687	85.71
7	MX02	MEXICO	372	176	33	390	589	345	613	85.46
8	CO01	COLOMBIA	320	147	28	401	672	401	677	85.36
9	GT08	GUATEMALA	374	171	31	399	637	399	660	85.29
10	GT05	GUATEMALA	369	171	30	390	583	390	625	85.2
11	CO16	COLOMBIA	367	172	30	399	590	399	590	85.14
12	ET02	ETHIOPIA	376	180	30	394	602	394	605	85.04
13	ET04	ETHIOPIA	367	176	30	396	544	396	550	85
14	CR10	COSTA RICA	361	183	31	396	608	396	619	84.92
15	ES01	EL SALVADOR	379	174	31	397	621	397	627	84.92
16	IN07	INDIA	383	181	29	390	593	390	612	84.92
17	NI04	NICARAGUA	385	172	28	390	628	367	634	84.75
18	CO05	COLOMBIA	378	180	28	397	541	397	548	84.71
19	CO06	COLOMBIA	394	180	31	401	582	401	588	84.71
20	CR09	COSTA RICA	372	176	29	392	612	392	613	84.71
21	CR03	COSTA RICA	378	174	30	396	579	396	591	84.68
22	CO11	COLOMBIA	392	178	29	408	560	408	568	84.61
23	NI08	NICARAGUA	370	167	30	390	647	390	652	84.61
24	GT06	GUATEMALA	379	172	27	392	641	392	656	84.57
25	CO12	COLOMBIA	374	174	31	403	602	403	604	84.54
26	CR02	COSTA RICA	369	176	34	399	546	399	569	84.54
27	CR08	COSTA RICA	365	174	28	403	560	403	564	84.54
28	CO10	COLOMBIA	392	172	28	405	616	405	626	84.5
29	IN13	INDIA	376	176	28	385	600	385	621	84.5
30	ES09	EL SALVADOR	392	178	29	410	588	410	594	84.46
31	ET03	ETHIOPIA	378	174	30	396	578	396	581	84.46
32	IN10	INDIA	376	174	30	388	618	388	618	84.46
33	GT01	GUATEMALA	356	162	30	394	644	394	651	84.4
34	ES08	EL SALVADOR	385	180	30	414	598	412	602	84.39
35	GT09	GUATEMALA	356	356	0	385	163	385	169	84.39
36	CR07	COSTA RICA	367	171	30	401	581	401	586	84.33

Appendix 2F: Coffee Green Parameters

A	B	C	D
CoffeeBeanParameterID	ParameterDescription	CoffeeBeanStage	Reading
1	Moisture	Green	7.83
1006	CountryCode	Green	013
1007	Region	Green	Brisas del Campanario/Intibuca
1008	Altitude	Green	1600
1009	Variety	Green	Catuai/Lempira
1010	Process	Green	Fully Washed
1011	ColorL	Green	46.78
1012	ColorA	Green	0.92
1013	ColorB	Green	13.21
1017	BeanSizeQ310	Green	4408
1018	BeanSizeQ350	Green	6251
1019	BeanSizeQ390	Green	7267
1020	BeanSizeMv3Mui	Green	6056

Appendix 2G: Coffee Roast Parameters

A	B	C	D
CoffeeBeanParameterID	ParameterDescription	CoffeeBeanStage	Reading
2	Moisture	Roast	0.21
1014	ColorL	Roast	27.44
1015	ColorA	Roast	4.75
1016	ColorB	Roast	4.61
1021	BeanSizeMv3Mui	Roast	7062
1022	BeanSizeQ310	Roast	5382
1023	BeanSizeQ350	Roast	7243
1024	BeanSizeQ390	Roast	8377

Appendix 2H: Coffee Roast Profile (Excerpt) Time / Temperature

A	B
Time (s) ▾	Value (FAHRENHEIT) ▾
0	415.4
1	415.2
2	414.7
3	413.6
4	411.1
5	407.1
6	401.9
7	396.3
8	389.5
9	382.8
10	375.4
11	368
12	360.7
13	353.5
14	346.3
15	339.2
16	332
17	325.4
18	318.7
19	312.2
20	306.1
21	300.2
22	294.1
23	288.5
24	282.9
25	277.7
26	272.5
27	267.4
28	262.6
29	257.9
30	253.7
31	249.2
32	244.9
33	241.1
34	237.2
35	233.6
36	230.2
37	226.6
38	223.3
39	220.1
40	217.2
41	214.1
42	211.3
43	208.6
44	206
45	203.5
46	201

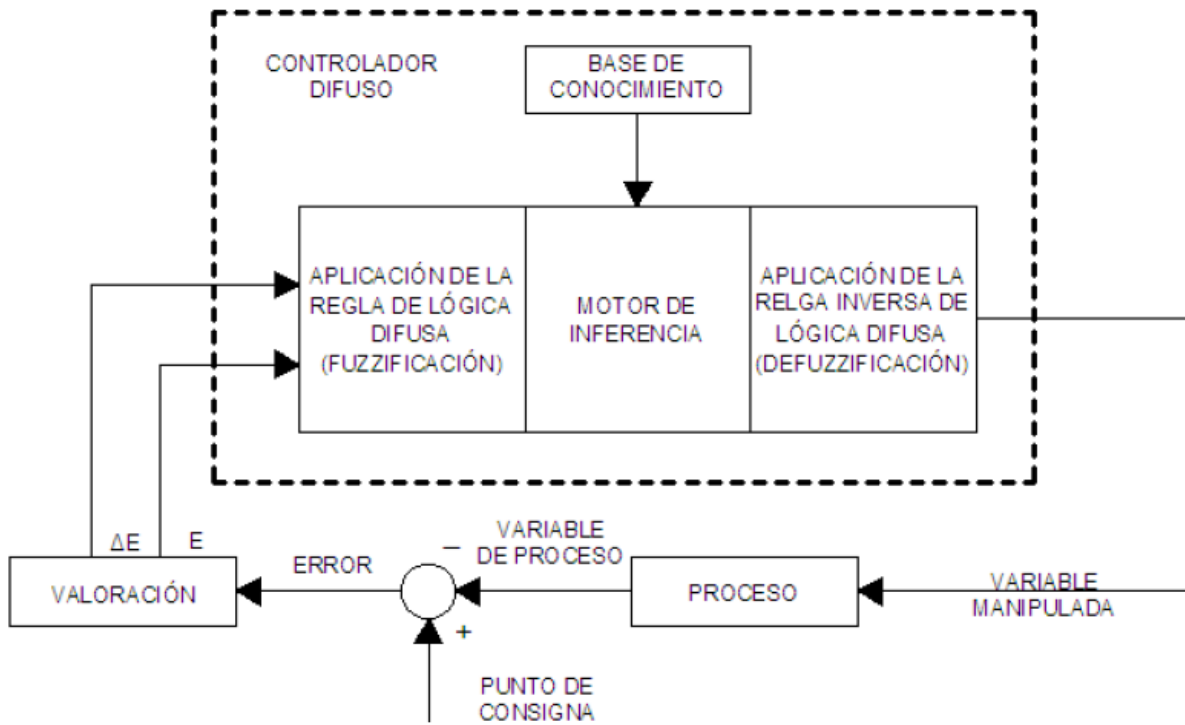
Appendix 2I: Coffee Roast Profile Time / Temperature Roasting Events

A	B	C	D
Time (s)	Temperature	Comment type	Comment
63	173.5	Gas	10
90	161.6	Turning point	
364	302.9	Gas	8
403	320.3	Comment	50
451	340	Gas	6
510	361.9	Gas	3
535	370	Comment	RD
606	390	First crack	
660	402.1	Gas	8
691	408.2	Gas	0

Appendix 2J: Drum Coffee Roaster Built at The Institute of Technology of Veracruz, Mexico



Appendix 2K: Integration of the Fuzzy Controller as Drum Temperature Regulator



Appendix 2L: Letter of Invitation for a Collaborative Effort Between the Coffee Technology Laboratory of the Food Research and Development of the Institute of Technology of Veracruz, Mexico



TECNOLÓGICO NACIONAL DE MÉXICO

Instituto Tecnológico de Veracruz

H. Veracruz, Ver. 29/octubre/2018

OFICIO No. ITV/D/1264/2018

Prof. Richard Chapman
Samuel Ginn College of Engineering
Department of Computer Science and Software Engineering (CSSE)
Auburn University, AL 36849

Hello,

I wish to submit to you the proposed activities of Mr. Javier Livio, your PhD student, that should perform during his stay at our Institute. The planned visiting time will be scheduled for the period from June to December 2019.

Five main activities are proposed:

- 1.- Act as a guest instructor in the field of Artificial Intelligence of the career of Computer Systems Engineering, emphasizing on Fuzzy Logic and its engineering applications.
- 2.- Serve as a key note speaker at our Multidisciplinary Research Congress addressing both system and applied fuzzy logic to students in electronic engineering
- 3.- To work here at our Institute, in a collaborative effort with other departments as electronic engineering, industrial engineering and sensory evaluation to collect the necessary data to produce the fuzzy controller
- 4.- Develop the algorithms of "Fuzzy Logic" to be applied on the automated coffee roaster.
- 5.- The publications resulting from the above collaboration will be shared with the participating parties from both, Auburn University and our Institute in Veracruz.

It is important to note that the advisor of the Mr. Livio will be Dr. Oscar González Ríos, head of the Coffee Technology Laboratory of the Food Research and Development Unit of this Institute, oscargr@itver.edu.mx; Tel. : + 52 229 9 34 57 01 ext. 203

Sincerely,

ATENTAMENTE
EXCELENCIA EN EDUCACIÓN TECNOLÓGICA*

ING. DAVID REYNIER VALDÉS
DIRECTOR

C.c.p.- Archivo.
DRV/OGR/lsc*



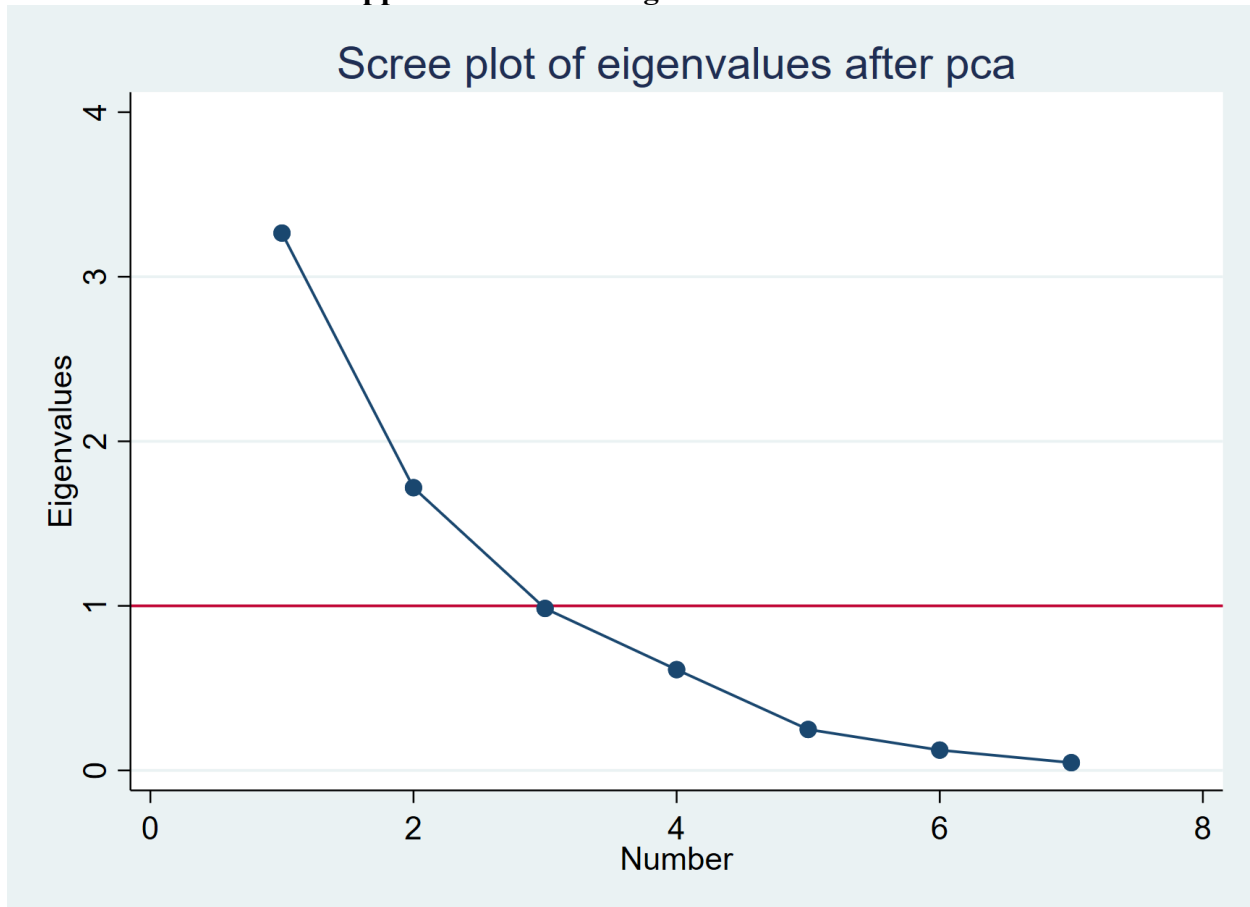
Calz. Miguel Ángel de Quevedo 2979, Cbl. Porfirio Díaz
C.P. 91897, Veracruz, Ver. Tel. (229) 934 1500
www.itver.edu.mx



Appendix 2M: Stata Principal Component Eigenvectors

Variable	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Unexplained
initialtemp	0.4610	-0.2879	0.0236	0.2215	-0.6449	0.4883	0.0266	0
mintemp	0.1581	0.1068	0.9260	0.2808	0.1552	-0.0542	0.0141	0
minsecs	0.3889	0.0483	-0.3676	0.7301	0.4054	-0.1154	-0.0255	0
maxtemp	0.4732	-0.2962	0.0143	-0.2398	-0.1950	-0.7691	-0.0331	0
maxsecs	0.3592	0.5482	-0.0313	-0.2427	-0.0119	0.1183	-0.7045	0
finaltemp	0.4246	-0.3257	0.0235	-0.4520	0.5837	0.3727	0.1705	0
seconds	0.2792	0.6395	-0.0709	-0.1355	-0.1297	-0.0303	0.6870	0

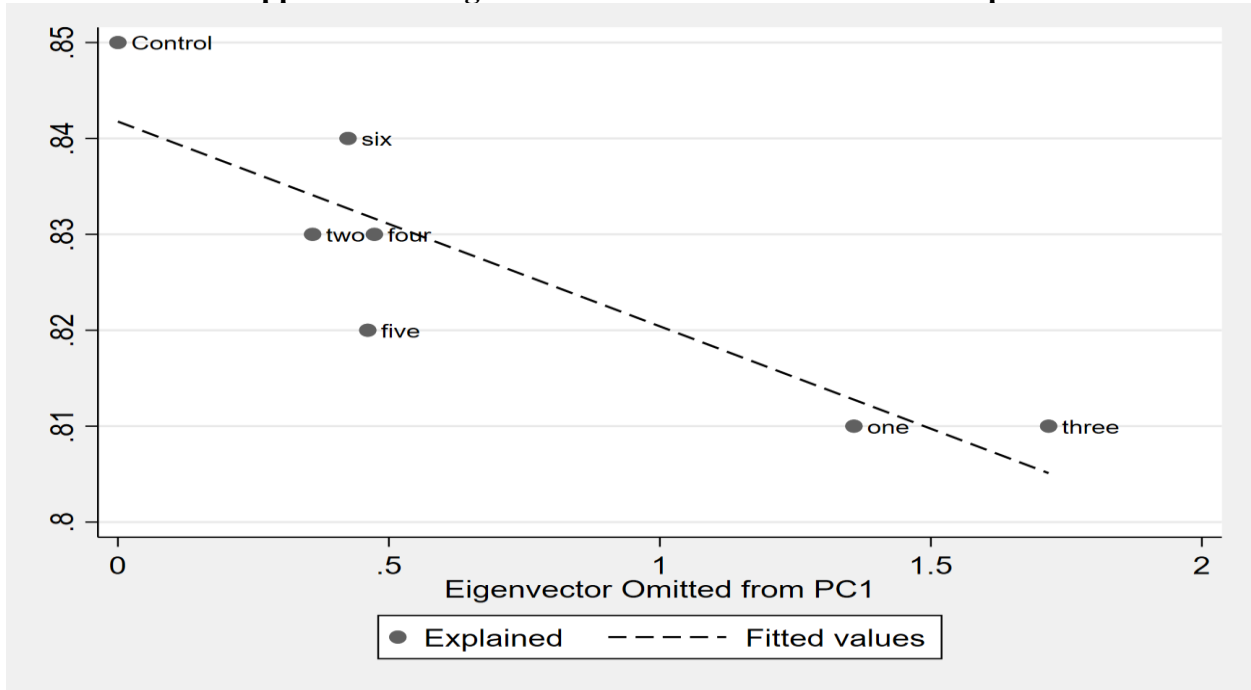
Appendix 2N: PCA Eigenvalues Scree Plot



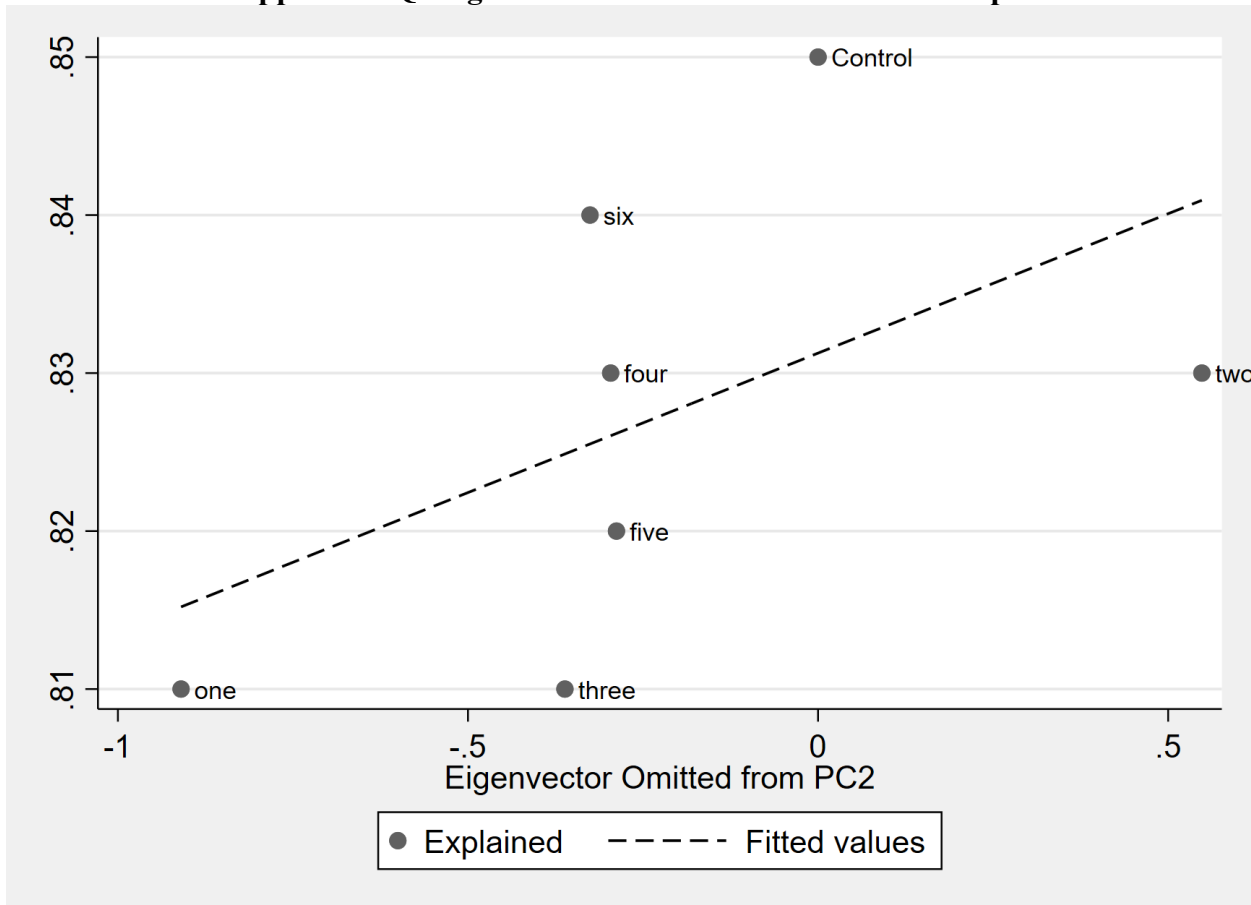
Appendix 20: Stata PCA Program (Script)

```
Do-file Editor - Principal Component Analysis v1.do
File Edit View Language Project Tools
Principal Component Analysis v1.do
1 * -----
2 * Principal Component Analysis of Roasting Profile impact on Final Quality of Coffee Beans
3 * Javier Livio July 28 2020
4 * -----
5 clear all
6 set more off
7
8 use C:\stata\data\Essay_2\CoffeeDataAllCountries_7-27-20.dta
9
10 global xlist initialtemp mintemp minsecs maxtemp maxsecs finaltemp seconds
11 * global id country
12 global ncomp 2
13
14 describe $xlist
15 summarize $xlist
16 corr $xlist
17
18 * Principal component analysis (PCA)
19 pca $xlist
20 * Scree plot of the eigenvalues
21 screeplot
22 screeplot, yline(1)
23
24 * Principal component analysis
25 pca $xlist, mineigen(1)
26 pca $xlist, comp($ncomp)
27 pca $xlist, comp($ncomp) blanks(.3)
28
29 * Component rotations
30 rotate, varimax
31 rotate, varimax blanks(.3)
32 rotate, clear
33
34 rotate, promax
35 rotate, promax blanks(.3)
36 rotate, clear
37
38 * Scatter plots of the loading and score variables
39 loadingplot
40 scoreplot
41 * scoreplot, mlabel($id)
42 scoreplot
43
44 * Loading/scores of the components
45 estat loadings
46 predict pc1 pc2, score
47
48 * KMO measure of sampling adequacy
49 estat kmo
50
```

Appendix 2P: Eigenvector Values not Included in Comp1



Appendix 2Q: Eigenvector Values not Included in Comp2



Appendix 3A: Updated IntelliTurk IRB



AUBURN UNIVERSITY INSTITUTIONAL REVIEW BOARD REQUEST for MODIFICATION

For information or help completing this form, contact: THE OFFICE OF RESEARCH COMPLIANCE (ORC)
Phone: 334-844-5966 E-Mail: IRBAdmin@auburn.edu Web Address: <http://www.auburn.edu/research/vpr/ohs>

In MS Word, click in the white boxes and type your text; double-click checkboxes to check/uncheck.

- Federal regulations require IRB approval before implementing proposed changes.
- Change means any change, in content or form, to the protocol, consent form, or any supportive materials (such as the Investigator's Brochure, questionnaires, surveys, advertisements, etc.). See Item 4 for more examples.
- Form must be populated using Adobe Acrobat / Pro 9 or greater standalone program (do not fill out in browser). Hand written forms will not be accepted.

1. Today's Date	4/23/2020
------------------------	-----------

2. Principal Investigator (PI)	
Principal Inves. (title): Marisha Speights Atkins Assistant Professor Department: CMDS Phone: 4-9634 AU E-mail: mls0096	Faculty PI (if PI is a student): Department: Phone: AU E-mail:
Contact person who should receive copies of IRB correspondence (Optional) Name: Phone: AU E-mail:	Department Head:

3. AU IRB Protocol Identification	
3.a. Protocol Number	20-146 EP 2003
3.b. Protocol Title	Crowdsourcing Intelligibility Determinations for Speech Related Disorders
3.c. Current Status of Protocol—For active studies, check ONE box at left; provide numbers and dates where applicable	
<input type="checkbox"/>	Study has not yet begun; no data has been entered collected
<input checked="" type="checkbox"/>	In progress If YES, number entered 4 Adverse events since last review 0
<input type="checkbox"/>	Data analysis only
Approval Dates: From 3/22/2020 To -----	
<input type="checkbox"/>	Funding Agency and Grant Number: AU Funding Information:
<input type="checkbox"/>	List any other institutions and/or IRBs associated with this project: <u>Auburn University # 17-042, #17-203, 19-311</u> <u>Children's Hospital of Atlanta CHOA00000450: 06-196</u>

4. Types of Change
Mark all that apply, and describe the changes in item 5
<input type="checkbox"/> Change Key Personnel Attach CITI forms for new personnel.

The Auburn University Institutional Review Board has approved this Document for use from
04/24/2020 to -----
 Protocol # 20-146 EX 2003

Appendix 3D: Testing the DNN Deployed Model with a Classification of a Spectrogram Image

```
Microsoft Visual Studio Debug Console
Predicting and Evaluation took: 14 seconds
Source=ImageClassificationTrainer; BinarySaver; Saving, Kind=Trace] Channel started
Source=ImageClassificationTrainer; EmptyDataView; Cursor, Kind=Trace] Channel started
Source=ImageClassificationTrainer; EmptyDataView; Cursor, Kind=Trace] Channel finished. Elapsed 00:00:00.0047374.
Source=ImageClassificationTrainer; EmptyDataView; Cursor, Kind=Trace] Channel disposed
Source=ImageClassificationTrainer; EmptyDataView; Cursor, Kind=Trace] Channel started
Source=ImageClassificationTrainer; EmptyDataView; Cursor, Kind=Trace] Channel finished. Elapsed 00:00:00.0014084.
Source=ImageClassificationTrainer; EmptyDataView; Cursor, Kind=Trace] Channel disposed
Source=ImageClassificationTrainer; BinarySaver; Write, Kind=Trace] Channel started
Source=ImageClassificationTrainer; BinarySaver; Write, Kind=Trace] Channel finished. Elapsed 00:00:00.0148332.
Source=ImageClassificationTrainer; BinarySaver; Write, Kind=Trace] Channel disposed
Source=ImageClassificationTrainer; BinarySaver; Saving, Kind=Trace] Channel finished. Elapsed 00:00:03.7518322.
Source=ImageClassificationTrainer; BinarySaver; Saving, Kind=Trace] Channel disposed
Model saved to: C:\code\DR-Speights\IntelliTurk.m\IntelliTurk.m\IntelliTurk.ML.Train.Model\bin\Debug\netcoreapp3.1\..\..\assets\outputs\dmeClassifier.zip
Source=ImageClassificationTrainer; SchemaBindableWrapper; Bind, Kind=Trace] Channel started
Source=ImageClassificationTrainer; SchemaBindableWrapper; Bind, Kind=Trace] Channel finished. Elapsed 00:00:00.0005190.
Source=ImageClassificationTrainer; SchemaBindableWrapper; Bind, Kind=Trace] Channel disposed
Source=ImageClassificationTrainer; MultiClassClassifierScore; GetEntireRow, Kind=Trace] Channel started
Source=ImageClassificationTrainer; MultiClassClassifierScore; GetEntireRow, Kind=Trace] Channel finished. Elapsed 00:00:00.0013054.
Source=ImageClassificationTrainer; MultiClassClassifierScore; GetEntireRow, Kind=Trace] Channel disposed
Source=ImageClassificationTrainer; SchemaBindableWrapper; Bind, Kind=Trace] Channel started
Source=ImageClassificationTrainer; SchemaBindableWrapper; Bind, Kind=Trace] Channel finished. Elapsed 00:00:00.0001575.
Source=ImageClassificationTrainer; SchemaBindableWrapper; Bind, Kind=Trace] Channel disposed
Spectrogram Image Filename : [luc101-15nm5-2-caap04_Bed_Easy.png], Scores : [0.061235804,0.41344717,0.22034149,0.23828746,0.06668812], Predicted DME Label : Easy
Press any key to finish

C:\code\DR-Speights\IntelliTurk.m\IntelliTurk.m\IntelliTurk.ML.Train.Model\bin\Debug\netcoreapp3.1\IntelliTurk.ML.Train.Model.exe (process 39136) exited with code 0.
To automatically close the console when debugging stops, enable Tools->Options->Debugging->Automatically close the console when debugging stops.
Press any key to close this window . . .
```