©Copyright 2023
Benjamin Ciaglo

# Developing a Framework for Metrics and Evaluation of the Impact of Acoustic-Prosodic Features in Synthesized Speech on Listener Perception in Dyadic Interactions

Benjamin Ciaglo

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2023

Committee:
Gina-Anne Levow
Richard Wright

Program Authorized to Offer Degree:
Department of Linguistics

University of Washington

## Abstract

Developing a Framework for Metrics and Evaluation of the Impact of Acoustic-Prosodic Features in Synthesized Speech on Listener Perception in Dyadic Interactions

Benjamin Ciaglo

Chair of the Supervisory Committee:
Gina-Anne Levow
Department of Linguistics

Acoustic-prosodic properties of conversational speech have been shown in prior research to impact the perceptions that listeners have towards the speakers in dyadic interactions. While correlations between the two have been consistently found, no standardized framework yet exists for systematically assessing the nature of such relationships in an easily reproducible fashion. The purpose of this thesis was to develop an experimental software prototype that facilitates researchers in their collection and assessment of data to explore the relationships between the audio qualities of a conversational dialogue agent's synthesized responses entrained on the acoustic-prosodic features of their human interlocutor's speech, and the way human listeners perceive that dialogue agent. Experiments were run to replicate results from past studies using the software developed.

# TABLE OF CONTENTS

# ACKNOWLEDGMENTS

First and foremost, I would like to sincerely thank my advisor, Gina-Anne Levow, who, throughout most of the course of my time in the CLMS program, not only guided me through the process of balancing the demands of completing this thesis alongside full-time engineering work, but who was also present and asking thoughtful questions exactly when they needed to be asked to point me further in the right direction, and expressing sincere interest in the development of the project over a sustained time period.

I would also like to thank Richard Wright, for introducing me to the concepts, processes, and analytical methods in experimental phonetics which were instrumental in delivering the software framework delivered in this work, and for sharing learning resources, experiences, and humor along the way.

The amount of love and support shown to me by my friends and family was also without doubt crucial in my ability to complete this work, including enthusiastic conversations on the possibilities of future iterations of this software, which helped me think outside the box about this project as it progressed. In particular, I would like to express profound gratitude to my mother, who encouraged me from the very moment I applied to this program, through to the end where she was the first person to try out the software delivered in this thesis in its earliest stage.

Finally, I would like to express deep appreciation to my wife, Jing, for all the times she listened to me talk through far too many programming problems just so that I could understand them myself.

# Chapter 1

# INTRODUCTION

This study puts forward the Perception Evaluation Framework[1], which integrates several existing repositories to create a standardized framework for data collection and analysis in experiments examining the impact on social perception of acoustic-prosodic entrainment in dyadic interactions. This study explores a proof-of-concept run of the software, wherein entrainment on pitch and volume was found to affect the perceived trustworthiness of a conversational agent. These results are analyzed and compared with the findings of previous studies, and the interdependence between these features is discussed and explored. The end output of the software is a clear recommendation for the kind of entrainment configuration a conversational agent should have in order to optimize the perception of trustworthiness users who interact with it have toward it. Areas for future improvement, and ethical considerations are explored.

Previous research has shown significant correlations between the acoustic-prosodic features of speech, and listeners' perceptions of a speaker. In both human-machine interactions, and human-human interactions, various types of entrainment, across multiple acoustic-prosodic features in synthesized spoken responses, measurably impact the perception of human listeners. While this impact on individual perceptual factors, like rapport or trustworthiness, has so far been shown in studies, no standardized framework yet exists for assessing such relationships between the suprasegmental qualities of speech in a dialogue, and the corresponding shift in the perception of a human participant in that dialogue. Without a dedicated experimental setup, this renders as unrealistic the comparison between how a given type of entrainment, for example, synchrony, of a given acoustic-prosodic feature, like pitch,

[1] https://github.com/benjamin-ciaglo/perception-evaluation-framework

alters a listener's perception of a speaker's character (i.e., both confidence, and competence simultaneously). Evaluating such multi-dimensional data requires trials, human participants, experimental setup, and controls to prevent externalities as well as the conversational flow itself from skewing results. A framework generalized such that it generates templated trials, and controls the semantic content of synthesized speech, will provide researchers with tools to understand more directly the relationship between the acoustic-prosodic delivery of that speech and how that delivery alters the intended outcome in terms of some given aspect of a listener's perception of the speaker. For example, a bank certainly wants a customer engaging with its spoken dialogue agents to come away with favorable ideas of the bank's reputation. While an automated salesperson is interacting with a customer, their company will prefer that customer to perceive the promotion of its products or services as persuasive. A developer will aim for the players of their video game to perceive its protagonist as heroic, rather than evil. Entraining acoustic-prosodic features in the synthesized voices of our interactive agents to achieve a particular outcome is feasible, yet identifying the combinations of which entrainment types to match with which acoustic-prosodic features to obtain that outcome requires rigorous analysis. With a defined, systematic process to create the suite of tests to gather the data needed for these analyses, as well as an interface to explore and understand that data, researchers will gain the capacity to focus on analyzing the nature of the interconnections in the relationships between entrainment types, prosodic features, and a given perceptual outcome, and developers will have a tool to optimize their system's synthesized voices toward a given perceptual outcome.

# Chapter 2

# RELATED WORK

## *2.1 Entrainment Impacts Social Perception and Behavior in Human-Human Interactions*

In the study, *Acoustic-Prosodic Entrainment and Social Behavior* (Levitan et al., 2012), participants played a cooperative game, and entrainment was extracted from their speech and measured on eight acoustic features. The degree of entrainment of these eight features was measured using manually labeled social variables produced by Amazon Mechanical Turk task workers who listened to dialogues from the Columbia Games Corpus. Objective measures were additionally used as assurance that annotators completed the labeling with integrity[1]. The researchers found that while female-male dyads entrain on all eight acoustical features, female-female and male-male pairs only entrain on a subset of these features, with male-male pairs not only entraining on fewer features, but also to a lesser degree. They also concluded that because the overall similarity between non-partner female-female speaking pairs is significantly larger than for non-partner female-male speaking pairs, the degree of similarity between female-female pairs is attributable to the overall similarity between females, rather than the effect of entrainment. The role that entrainment plays also differs between the sex/gender composition of speaking pairs. It was found to be more important to the perception of social behavior in dialogues between female-male speaking pairs, and more important to the "smoothness and flow" in dialogues between male-male speaking pairs. Again, since this study conflated sex with gender, it is unclear whether the role played by entrainment

[1]It's important to note that this paper conflated sex with gender, discussing both with the implicit assumption that gender is binary. Since this viewpoint is unscientific for a variety of reasons, this is a disclaimer that such views implied below are not shared by the author of this thesis; they are a summary of findings as reported in past research.

is dependent on a speaker's gender, their sex, a complex interplay between the two, or, in fact, none of the above but rather associated confounding factors. Lastly, when intensity or speaking rate is especially high or low, the study found that entrainment becomes more pronounced.

Researchers found that convergence (as the form of entrainment) in mean syllable duration is correlated with cooperation in the decision-making process during a Prisoner Dilemma game in the study, *Convergence of speech rate in conversation predicts cooperation* (Manson et al., 2013). They measured dyadic convergence on fundamental frequency, variation in fundamental frequency, and mean syllable duration, and their correlations with cooperation in the same game. Several measures of coordinated laughter, and laughter/speech coordination, as well as a language-matching score (LSM), were also recorded. Dyads showing greater coordinated laughter, greater verbal convergence (higher LSM score), and greater vocal convergence (only significantly with mean syllable duration) were more likely to cooperate during the game; these factors predicted co-participants' ratings of warmth and competence. Participants were found more likely to cooperate if they grew up in a wealthier area, and toward co-participants they found to have more attractive facial features. Participants were less likely to cooperate if they scored higher on a primary psychopathy measure toward co-participants who interrupted them more frequently, and toward co-participants with whom they found "no reliable cues to future interaction" (or "common ground").

These studies provide evidence that not only do humans entrain on acoustic-prosodic features naturally in human-human dyadic interactions, but also that such entrainment has noticeable impacts on social behavior that differs by context. Such findings provide grounds in interpreting analytical methods developed alongside the software delivered in this thesis to be also applicable to speech in human-human interactions. They also provide grounds for the necessity of a standardized framework to achieve reproducible results across contexts.

### *2.2 Entrainment Impacts Social Perception and Behavior Across Languages*

*Contrasting Multi-Lingual Prosodic Cues to Predict Verbal Feedback for Rapport* (Wang & Levow, 2011) provides a feature analysis that was conducted across multiple languages, including Arabic, English, and Spanish. Results were reported with weight-based rankings for features assigned to each language. These rankings highlight which prosodic features were used extensively for speaker cues to elicit listener verbal feedback in each language; they also highlight which features are used in each language distinctively. In Arabic, the researchers found that pitch is used both extensively and distinctively in cuing verbal feedback, in words both preceding and following pausal intervals (intervals where no speech occurs). In English and Spanish, both pitch and intensity are utilized by the speaker to cue verbal feedback, while vocalic and pause duration are used exclusively in Spanish.

In *Prosodic entrainment and trust in human-computer interaction* (Benus et al., 2018), researchers investigated the correlation between the trust of a human participant in whether an avatar can provide good advice, and several acoustic-prosodic features (including pitch, intensity, speech rate, and voice quality), across several forms of entrainment (including similarity, convergence, and synchrony). All participants were native speakers of Slovak. They found that a user's sex influences the relationship between their trust in an avatar and the amount that avatar's speech entrains on theirs, that the advice of avatars whose speech disentrains is preferred by females, and that the level of entrainment adapted in synthesized speech needs to be amplified to affect the trust of humans toward computer avatars (to account for low variability of features). Increases in both ASR accuracy, and in learning were observed where entrainment on pitch and intensity occurred. Entrainment and interactions perceived positively were found to be correlated. While researchers stated no consensus regarding the role of biological sex in relation to the degree and function of entrainment, more entrainment was observed between mixed sex dyads. Openness to experience, as a personality trait measured via sociometric questionnaire, was also found to be correlated to increases in acoustic-prosodic entrainment. It is also noted by the researchers that the relationship

between trust and entrainment can be due to cultural differences between native speakers of different languages, as well as linguistic differences between the languages themselves.

Given the above, there is evidence that the social impacts of acoustic-prosodic entrainment occur *across languages*, in human-human interactions as well as in human-computer interactions. It is also clear that the nature of acoustic-prosodic entrainment, and its impacts, differ from one language to the next. This implies that the utility of the software developed in this thesis extends beyond data collection and analysis in any one particular language.

### *2.3 Entrainment in Learning Systems*

In *Prosodic Entrainment and Tutoring Dialogue Success* (Thomason et al., 2013), researchers found that learning gain is positively correlated with entrainment for several pitch features for all students interacting with a tutoring dialogue system, and significantly for students with high pre-test learning scores. Loudness min was found to be significantly correlated positively for students with low pre-test scores. On loudness min/max features, male mean entrainment was reported to be higher than female mean entrainment.

In the study, *Naturalness and rapport in a pitch adaptive learning companion* (Lubold et al., 2015), researchers applied pitch adaptation using three different methods based on analysis of human-human entrainment, to measure the relationship between perceptions of rapport and pitch adaptation in speech. Data was collected from four individuals who interacted with a learning companion employing each form of pitch adaptation, and analysis was crowd-sourced using Amazon Mechanical Turk. The three adaptations were compared to a baseline TTS on perceptions of rapport and naturalness reported by third-party observers. The measured form of entrainment on pitch in this case was proximity. Several styles of proximal entrainment on pitch are compared, including two where the speaker's utterance was resized to the length of the proposed TTS output before applying the utterance's contour to adapt the pitch of the synthesized response either to the contour of the speaker's utterance, or the same contour shifted based on the pitch of the dialogue agent. This both

maximized the level of entrainment and controlled for differences in utterance length. Perceptual effects of proximal entrainment on pitch were measured by asking AMT workers to evaluate adjacent turns between human speakers, and the responses of dialogue agents. Findings included that adapting a learning companion's pitch to that of its human partner does affect naturalness and rapport.

There are domains in which the benefits of understanding the relationships between entrainment and social behavior are clear. Learning is one such domain, where a direct material outcome is associated with programming a conversational dialogue agent to respond with a particular entrainment profile. This further supports the need for a standardized framework to understand such acoustic-prosodic interactions, especially given that achieving those material outcomes will require different entrainment profiles across contexts and languages.

### *2.4 Entrainment's Effects are Not Necessarily Opposite of Disentrainment's Effects*

Two novel measures of synchrony were used in *Disentrainment may be a positive thing: A novel measure of unsigned acoustic-prosodic synchrony, and its relation to speaker engagement* (Pérez et al., 2016), in which researchers found that the development of conversation may be positively affected by disentrainment. Of these two measures of synchrony, one penalizes disentrainment, while the other rewards it, though both reward entrainment. The Columbia Games Corpus was used for acoustic-prosodic analysis, and annotators were asked a series of questions to measure third-party perception of the speakers' exchanges. On the second measure, where both entrainment and disentrainment were rewarded, statistically significant correlations were found between positive perceptions of dialogue, and prosodic entrainment between speakers. Notably, the paper distinguishes between neither sex, nor gender.

Researchers in *An empirical study of the effect of acoustic-prosodic entrainment on the perceived trustworthiness of conversational avatars* (Gálvez, Gravano, et al., 2020) measured how entrainment affects users' perception of a system's trustworthiness, and ability. They

found the acoustic-prosodic features do not behave independently of the way other acoustic-prosodic features behave. For example, when positive correlations between speech intensity and perceptions were observed, simultaneously disentraining on speech rate was found to negate the effects. It was also found that entraining on speech rate by itself did not improve user perceptions; this implies the relationships between individual acoustic-prosodic features are multi-dimensional, and interdependent. Following from these last two points, at the policy level, altering the level of entrainment of one feature may have no effect, though it does at the exposure level.

The above studies offer strong evidence that the nature of entrainment's impacts on social perception and behavior is multi-faceted and complex. To achieve the opposite outcome in a sophisticated learning system, for example, it is not sufficient to assume you would use the opposite entrainment profile for a given acoustic-prosodic feature. This further suggests that a standardized framework is needed to systematically analyze the multidimensional relationships between such features, and the impacts they have on perception.

### *2.5 Techniques*

Researchers found that entrainment is associated with both likability and task success in the study, *Implementing acoustic-prosodic entrainment in a conversational avatar* (Levitan et al., 2016). Their evaluation method shows statistical significance between speech synthesized using their algorithm, and target outputs determined by human-human interactions.

In *Measuring Acoustic-Prosodic Entrainment with Respect to Multiple Levels and Dimensions* (Levitan & Hirschberg, 2011), researchers examine four acoustic-prosodic features, including intensity, pitch, voice quality, and speaking rate, in the Columbia Games Corpus. At the session level, they measure proximity and convergence, and at the turn level, they measure proximity, convergence, and synchrony. Results show evidence for entrainment across all four features in natural human-to-human dialogues; at both the session and the turn levels, intensity is found the most consistently, followed by pitch, then vocal quality and speaking rate.

These studies were useful in providing either direct or indirect inspiration for the techniques implemented in this project to generate synthesized speech entrained on a particular acoustic-prosodic feature, or to verify that entrainment occurred.

# Chapter 3

# METHODOLOGY

This chapter contains an overview of the methodology used across the various components of the system, including algorithms, data structures, architectures, and frameworks. Data collection and recruitment procedures are also described, followed by a description of the system's final output.

## *3.1 System Overview*

The motivation behind building the Perception Evaluation Framework was to provide an environment in which researchers can run reproducible experiments testing the impacts of various acoustic-prosodic features of a conversational agent's speech on the social perception of human listeners to that agent's interactions. The goal is to create an end-to-end pipeline, from the data collection phase all the way to the analysis phase, that facilitates systematic examination of such relationships. The input to the system is the perceptual factor the researcher would like to understand in terms of its relationship with acoustic-prosodic entrainment, and the output is an optimal entrainment profile that can be used to inform the programming or configuring of a conversational avatar such that it exhibits as much of that perceptual factor as possible. The repository combines and modifies several existing repos, including a Python wrapper called PyWORLD (JeremyCCHsu, 2023) which is built around the WORLD speech analysis, manipulation, and synthesis system (Morise, 2016a; Morise et al., 2016b), in order to access the Harvest f0 method introduced in (Morise 2017), a Python wrapper to call Amazon Polly from AWS documentation (awsdocs, 2021), and the Speak Tool introduced in (Song, et. al. 2022).

## 3.2 System Flow

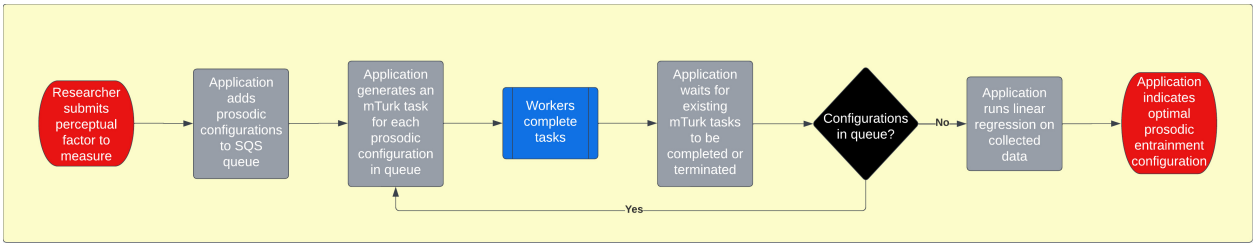


Figure 3.1: System flow diagram.

The first thing a researcher needs to update before running the software, are variables which specify the perceptual factor which will be displayed to participants in the study that will be launched. When the user kicks off an experimental run, an SQS queue is populated with acoustic-prosodic configurations which will be used to decide how to entrain the conversational agent's response to a given participant. The application then generates an mTurk task for each configuration on the SQS queue. During that task, workers are asked to upload a recording where they respond to a content-neutral question, seen in 3.2, then listen to a synthesized response entrained on their input speech and answer a question regarding their impression of it, seen in 3.3. As workers complete tasks, the queue is gradually cleared; when the researcher finds that it is empty, they can decide whether the data is high enough quality to move on to the next stage of data analysis, or to submit another round of entrainment configurations for additional data collection. In the end, the application outputs the entrainment configuration a conversational agent needs to adapt in response to a human interlocutor in order to optimize listeners' perception of the factor input into the software by the researcher for examination.

## 3.3 Input Parameters

The initial input to the system should be a trait perceivable by an observer of a dyadic interaction. While prompts describe how the input parameter will be used, discretion is left

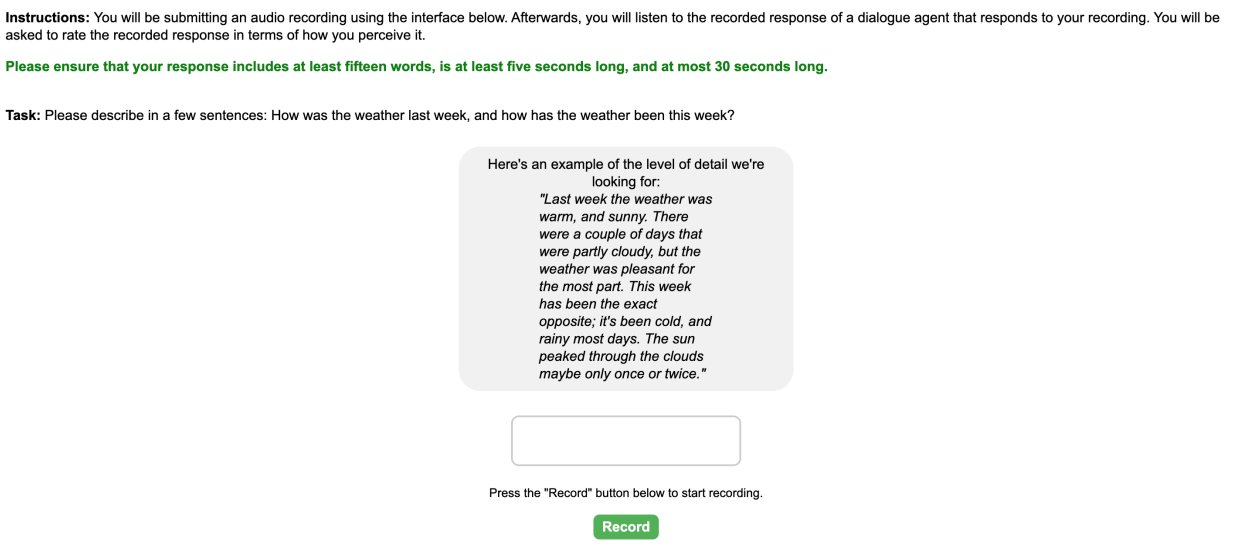


Figure 3.2: First screen, where user uploads recording; derived from (Song, et. al. 2022).

to users of the system to understand what such traits are appropriate inputs. The survey question in the second step of data collection is presented in such a way as to allow users to fill in the blank, to capture any inflection or nuance they may deem necessary given the context. In this way, users see the context surrounding the question that will measure the trait they chose to input. The procedure used to produce the results in this paper are discussed in greater detail in Section 3.4.3. Currently, the number of tasks for each entrainment configuration is hardcoded in the mTurk module, and configured manually if using Appen's interface instead. For the purposes of the proof-of-concept results presented in this paper, the target demographics were limited to participants in the United States who are 18 and older, fluent English speakers, any gender, who can speak, read, and listen to audio, and who have a microphone.

## *3.4 Configuration Management*

Once the user has input information regarding the question to ask participants on mTurk, and which personality or perceptual trait that question will measure, the application then asks

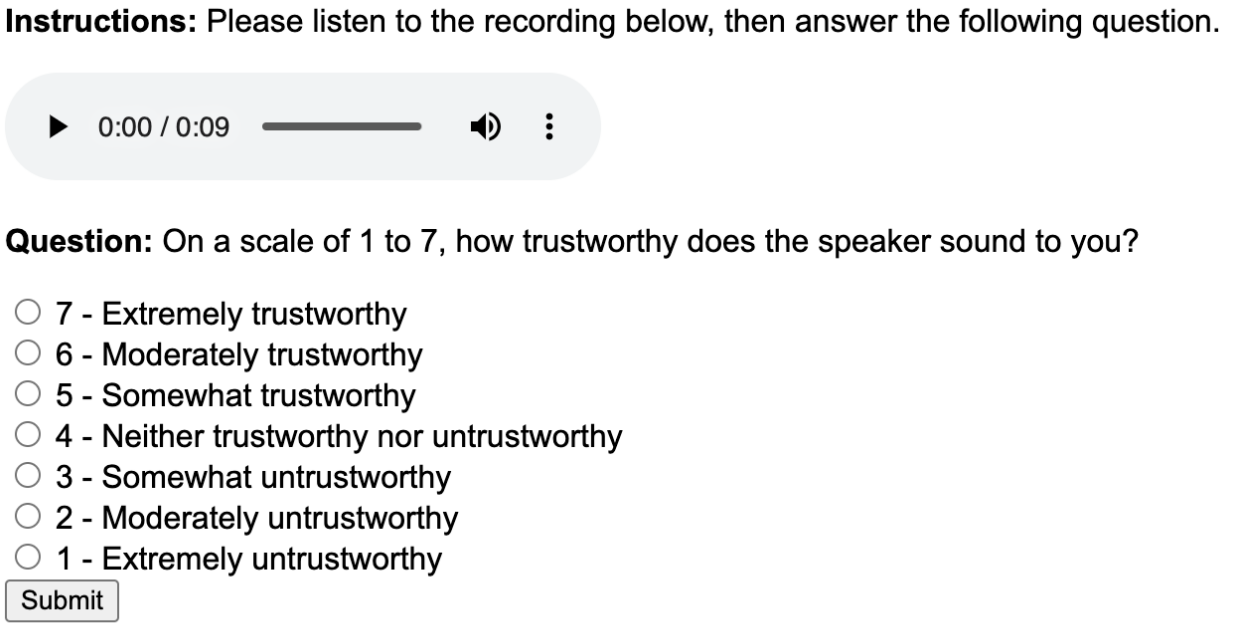


Figure 3.3: Second screen, where user listens to synthesized response of dialogue agent, and answers a question regarding how they perceive it.

the user to select which types of entrainment should be measured, and for which acoustic-prosodic features. The combinations of features and entrainment types that will be measured are numerous, as shown in the Task Groups described in Table 3.1 below. Each combination, which from here will be referred to as an entrainment *configuration*, will be denoted in an item pushed to an SQS queue in AWS.

### *3.5 Task Generation*

For each entrainment configuration item pushed onto the SQS queue described in Section 3.2, a task is generated in either mTurk or Appen to ensure that a participant is presented with a synthesized voice response entrained with that configuration, along with a survey question specifying the perceptual trait the user input. As shown in Table 3.1, if all acoustic-prosodic

Table 3.1: Task groups along with columns describing how many features are entrained for that group, and the corresponding number of combinations comprising each group.

| Task Group | Entrain | Disentrain | None | Combinations |
|---|---|---|---|---|
| A | F1, F2, F3 | - | - | 27 |
| B | F1, F2 | F3 | - | 27 |
| C | F1, F2 | - | F3 | 27 |
| D | F1 | F2, F3 | - | 27 |
| E | F1 | F2 | F3 | 27 |
| F | F1 | - | F2, F3 | 9 |
| G | - | F1, F2, F3 | - | 27 |
| H | - | F1, F2 | F3 | 27 |
| I | - | F1 | F2, F3 | 9 |
| J | - | - | F1, F2, F3 | 1 |

features are selected, as well as all entrainment types, the number of tasks that will be run is 208. For the purpose of the proof-of-concept results produced in this paper, the number of tasks will be reduced by focusing instead on two entrainment features, pitch and volume, each associated with an entrainment type, synchrony and convergence. Therefore, the number of configurations that will be generated are as shown in Table 3.2 below, yielding a total of 9. For the results to be generated in this thesis, three tasks will be run for each configuration to ensure some degree of generalizability, leading to a total of 27 tasks.

## *3.6 Task Completion*

### *3.6.1 Recruitment and Consent Procedures*

Note that the contents of the recruitment and consent pages are hardcoded in the application currently. Tasks are either posted as jobs on mTurk by the application, or configured on

Table 3.2: Task groups and the corresponding number of combinations comprising each group if measuring only two entrainment features, each with its own entrainment type (pitch synchrony, volume convergence).

| Task Group | Entrain | Disentrain | None | Combinations |
|---|---|---|---|---|
| A | F1, F2 | - | - | 1 |
| B | F1 | F2 | - | 1 |
| C | F1 | - | F2 | 1 |
| D | - | F1, F2 | - | 1 |
| E | - | F1 | F2 | 1 |
| F | F2 | F1 | - | 1 |
| G | F2 | - | F1 | 1 |
| H | - | F2 | F1 | 1 |
| I | - | - | F1, F2 | 1 |

an alternate platform of the user's choice. The recruitment page (which is the first page of the site) contains much of the same information as the consent page. Before beginning participation, subjects are shown a consent page describing the study's purpose, and its minimal potential risks. In order to begin the study, this page must first be read, and consent confirmed. On this page, they are informed they have the right to refuse participation or withdraw from the study at any time, and that they will be given a payment pro-rated for the tasks they completed if they do. The schedule/timing of the payment is described, as well as the purpose of the payment. The page specifies that if a participant is deemed to be taking an unreasonable amount of time compared to the estimated amount of time for the current task, they may be removed from the study. Contact information of the PI for the sake of answering clarifying questions they may have is provided, and the estimated amount of time for participation is specified, as likely fifteen minutes or less, along with the number of pages total in the task. The participant is reminded to print the consent page if they

would like a record, and informed there are no plans to return to them the results of the research. Lastly, they are asked to confirm they have a microphone, consent to having their voice recorded, are in the United States, can speak, hear, and write English, are 18 years of age or older, are not a UW employee, family member, or student, and that they understand a small set of comprehension questions displayed on the page.

### *3.6.2 Audio Data Collection, Speech Analysis, and Entrainment*

When the participant begins the study, they are prompted with a written question: "Please describe in a few sentences: How was the weather last week, and how has the weather been this week?" The participant then is instructed to upload a spoken response to the question. When they do so, the web page uploads their recording to the server, which then triggers a function to analyze and extract audio features from the recording using the Harvest f0 method introduced in (Morise 2017), to determine pitch values at a particular time segment. These values are represented as an array where the values are sequential by time.

The approach used to actually entrain the dialogue agent is intended to approximate the procedure in (Levitan & Hirschberg, 2011) used to measure pitch synchrony. In that work, it is described as the Pearson $r$ between two adjacent IPUs in a dyadic interaction. Though not formally stated in their work, this was interpreted as:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{3.1}$$

The entrainment approach used to achieve volume convergence is somewhat simpler, and was inspired by the (Levitan & Hirschberg, 2011) approach to measuring session-level convergence, instead adapting it to the turn-level. The mean intensity of the first half of the participant's recording is calculated, followed by the mean intensity of only its second half. The dialogue agent's response is synthesized by taking the difference between the mean of the two halves, then setting the intensity of the beginning of the synthesized response to a value equal to the value of the second half minus that difference. A gradient is used so that

the convergence is gradual; disentrainment simply inverts the procedure above so that the mean intensity diverges in the second half:

$$\text{gradient} = \{x_i\}_{i=0}^{n-1} = \left\{ s + i \cdot \frac{e-s}{n-1} \mid i \in \{0, 1, 2, \ldots, n-1\} \right\} \tag{3.2}$$

### *3.6.3 Perception Data Collection*

Once the audio data has been analyzed, it is then used to produce a synthesized response to display to the participant. That response is saved to the server, so that it is available for the participant to access by means of a link made available on the task page. The next step is for the participant to click the link to listen to responses served to the user. The amount of time is minimized between the uploading of their spoken recording, and their listening to the entrained, synthesized response. Despite this, there is a short delay. The effects of such delay were not found during literature review; regardless, a future improvement to this framework would be to facilitate such dialogue in real-time. They will be asked, "On a scale of 1 to 7, how _ does the speaker sound to you?" For the purpose of the proof-of-concept in this paper, trustworthiness will be the perceptual trait measured.

## *3.7 Task Management*

Once a generated task has begun, its corresponding configuration is removed from the queue described in Section 3.2. If the data for the task is incomplete for whatever reason, its corresponding entrainment configuration is not placed back on the queue; currently, incomplete tasks require manual intervention of the researcher to check the data collected, then launch more tasks if they discover the data that has currently been collected is insufficient.

## *3.8 Correlating Features with Perception*

This step is currently not automated, and was performed manually using R against the data collected, for the proof-of-concept. Once all data has been collected in an experimental run, p-values are obtained using Likelihood Ratio Tests run with ANOVA analysis of full and

reduced model, inspired by (Lubold et al., 2015) and (Winter, 2013). A mixed model is employed, where each entrainment feature of a given type is a fixed effect, and a different intercept is assumed for each subject to account for random between-subject variation. The dependent variable, then, is the perceptual trait under measurement. Likert scores for perceptual traits are transformed from ordinal values to continuous values using a logit transform, a common method for transformation of ordinal data to render it in continuous values (Armitage & Berry, 1994). First, interdependence between the entrainment configurations is tested by using ANOVA on a model with interactions between configurations versus a model with fixed effects assumed independent. If the entrainment configurations are found to be interdependent, then an intercept-only model is compared with a model containing the entrainment configurations. If they are found not to be interdependent, then a model with both fixed effects is compared with a model with only one fixed effect, then to a model with only the other. Note, that in its current form, this software is setup to run by testing only two feature-type configurations at a time, for the sake of proof-of-concept.

# Chapter 4

# EXPERIMENTS & ANALYSIS

## *4.1 Data Collection*

Data was collected by deploying jobs on the Appen platform, and deploying the software suite to an AWS EC2 instance. Nine entrainment configurations were deployed to the SQS queue, including all combinations of entrainment/disentrainment on pitch, and entrainment/disentrainment on volume, including the possibility of neither entrainment nor disentrainment on either of the two features. The *type* of entrainment on the pitch feature was synchrony, while the *type* of entrainment on the volume feature was convergence. For each configuration, three data points were collected, leading to a total of 27 data points.

## *4.2 Deployment Architecture and Setup*

A domain was purchased, and its A record was pointed to the EC2 instance on which the software suite was deployed to run on an Nginx server. Webpages were served over the instance's UNIX socket, and permissions were configured such that the instance could send and receive requests to and from an SQS queue, as well as from the Amazon Polly API.

## *4.3 Platform Considerations*

At the time of this writing, onboarding of new requesters to mTurk's payment system was somewhat complicated by the requirements to onboard through Amazon Payments, which does not allow one to prepay for work completed on the platform. The alternative was to wait for the platform to approve paying for a certain amount of work ahead of time based on previous months' AWS bills. Because the project account had been created less than one month prior to being ready-to-deploy the experiments, this meant another option had

to be pursued; in this case, that alternative was Appen. A core benefit of the platform is that it leverages a managed crowd, such that contributors found to be attempting fraud are removed from the platform. Theirs is a data collection platform typically used for collection of ML training data rather than for surveys. This led to various challenges when beginning the data collection process, in terms of synchronizing the Appen jobs with the survey on the external server. The primary issue was that participants would find ways to mark their Appen job as complete, without having done any work on the external server; they were paid, yet no data was successfully collected. Initially, this was due to having launched the job with a hardcoded completion code, which was readable by inspecting the Appen job page's rendered HTML. On the next attempt, a regex was hardcoded for validation into the completion code's field. Clever participants did the same thing: entered a matching code without having completed any work. Finally, the site's quality control mechanism (Appen Help Center, n.d.) was employed: test questions. Multiple questions were added to each job, asking for multiple completion codes that could only be found by reaching the end of the study on the external server. No validation was hardcoded into the HTML; rather, test questions were used as filters such that participants who failed to enter the correct completion codes during Test Mode (usually such responses were outdated completion codes from previous job batches, or spammed URLs to external advertisements) were not subsequently allowed to input completion codes in Work Mode. Participants who correctly answered the questions in Test Mode were allowed to re-enter the same completion codes in Work Mode in order to receive payment through Appen's payment system.

## *4.4 Data Analysis*

### *4.4.1 Data Description*

This subsection describes overall statistics surrounding the experiments run. Specific datapoints were discarded for a variety of reasons, as described in the *Inclusion and Exclusion Criteria* section below, and listed in A.2 in Appendix A. Throughout the course of experi-

mental data collection, a total of 77 data points were collected from a total of 15 participants. 7 total data points were collected without entrainment on any features, 13 total data points were collected where entrainment only on pitch was attempted, 10 total data points where disentrainment only on pitch was attempted, 5 total data points where entrainment only on volume was attempted, 5 total data points where disentrainment only on volume was attempted, 8 total data points where entrainment on both pitch and volume was attempted, 9 total data points where disentrainment on both pitch and volume was attempted, 12 total datapoints where entrainment on pitch was attempted and disentrainment on volume was attempted, and 8 total data points where disentrainment on pitch was attempted and entrainment on volume was attempted. Of the data points remaining, 27 data points were analyzed, collected from eleven unique participants, as shown in A.1 in Appendix A. For each entrainment configuration (condition), three data points were analyzed.

For the sake of respecting privacy, and for proving the utility of the software for studies wishing to retain IRB exempt status, alphanumeric codes were stored prior to analysis, rather than worker IDs, and the two were not linked. Despite efforts to prevent the same subject from repeating the experiment multiple times, judging from timestamps, and reviewing the worker recordings, it seems likely that some subjects had multiple distinct worker IDs, and completed the experiment several times. Because some subjects tended to complete the recordings with less problematic audio, such subjects' data points were kept rather than discarded. Listening to the participant's recordings, all participants tended to have accents likely of Hispanic origin, indicating they were not L1 English speakers. Of the eleven participants who contributed 27 total retained data points, ten subjects had masculine-presenting voices (judging from listening to the recordings), and one subject had a feminine presenting voice (who contributed one data point). One subject was particularly active and reliable, contributing a total of six data points. However, the *mode* of data points contributed for a given subject was one, the *median* was 2, and the *mean* was 2.5.

### *4.4.2 Inclusion and Exclusion Criteria*

Exactly fifty data points were discarded for a variety of reasons, all of which can be grouped under the following broader categories:

Audio Quality Issues: These included issues like clipping, and choppy audio. These issues affected 22 of the discarded datapoints.

Dropout: This issue denotes random fluctuations in volume, likely caused by microphone movement, network failure, or other audio issues where the participant's voice became temporarily inaudible. This issue affected 10 of the discarded datapoints.

Background noise: Background noise was common, though louder in some cases than others. These often included high-pitched humming, and in some cases, animal sounds, or other people talking. These issues affected 13 of the discarded data points.

Correlation Issues: This denotes issues entraining the conversational dialogue agent to the acoustic-prosodic features in the participant's speech. Notably, this applies to data points where the condition only involved entrainment or disentrainment on pitch, as this level of validation is not implemented for other features/combinations yet. This issue affected nine of the discarded data points.

Silence: This is self-explanatory; the participant's voice and environment was totally inaudible. This affected one data point.

Muffled Sound: This describes scenarios where the participant's voice was audible, though clearly muffled, likely due to the orientation of their microphone. This affected 11 data points.

### *4.4.3 Statistical Tests*

Likelihood ratio tests were performed on the data with ANOVA analysis (Winter, 2013) as inspired by research using ANOVA on similar data (Lubold et. al. 2015). Mixed models were employed where each entrainment feature was treated as a fixed effect, and a different intercept was assumed for each subject. Likert scores were made continuous using a logit transform, where they were first divided by seven to convert them to proportions of

the maximum score. Then the logit function from the car package (R Core Team 2019) in R was used to perform the transformation, where the proportions were remapped to (0.025, 0.975). Entrainment features were treated as categorical variables. For example, the value for entrainment on pitch was treated as a 1, the value for disentrainment on pitch was treated as a -1, and the value for the absence of entrainment was treated as a 0. The reasoning for this was, as found in (Gálvez 2020), for example, that the effect of disentrainment on pitch will not necessarily be the opposite of the effect of entrainment on pitch; meaning that treating them similarly to the treatment of categoricals in (Winter, 2013) is more appropriate than to treat them as on a linear scale alongside the Likert scores for trustworthiness. From a high level, and without conducting such analysis, no particular

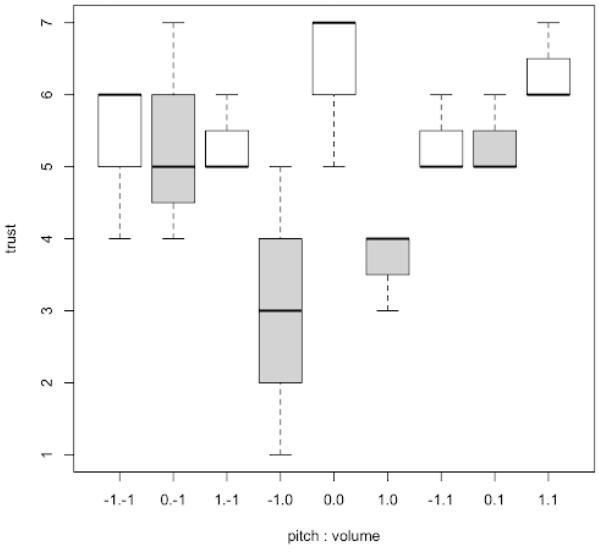


Figure 4.1: Configurations and their corresponding trustworthiness ratings.

pattern is obvious. The one exception to this, is that it seems to be the case that entraining or disentraining on pitch at all has a negative impact on the trustworthiness of a speaker, as seen in Figure 4.1. To analyze the data on a deeper level, first, interdependence was tested by comparing a full model including interactions, with a reduced model that treated

the entrainment factors as independent from one another. Interactions between pitch and volume were found to be significant ($\chi^2(1)$=10.797, p=0.02894). Then, a full model with interactions was compared with an intercept-only model, containing only the per-subject intercepts and a constant. This, combined with the summary of the full model, revealed that

Table 4.1: The intercept of the model was estimated at 2.7305 with an error of 0.5015, and high confidence (t-value of 5.445). Note that disentrainment on volume was not found to significantly impact perception of trustworthiness and so was excluded from this table.

| Configuration | Entrain | Disentrain | None | Likert Log Odds | Std Errs |
|---|---|---|---|---|---|
| 1 | - | Pitch | Volume | -3.0844 | 0.7092 |
| 2 | Pitch | - | Volume | -2.6395 | 0.7092 |
| 3 | Volume | - | Pitch | -1.6033 | 0.7092 |
| 4 | - | Pitch, Volume | - | 2.6770 | 2.669 |
| 5 | Pitch | Volume | - | 2.1663 | 1.0029 |
| 6 | Volume | Pitch | - | 3.0844 | 1.0029 |
| 7 | Pitch, Volume | - | - | 3.8354 | 1.0029 |

both entrainment and disentrainment on the interdependent features of pitch and volume affected perception of trustworthiness ($\chi^2(1)$=18.317, p=0.01897) as in Table 4.1. When comparing the reduced model with a model containing only the effects of volume, we see a significant impact ($\chi^2(1)$=6.7885, p=0.03357), as well as when comparing an intercept-only model with a model containing only the effects of pitch, where we see also see an effect of pitch disentrainment on perception of trustworthiness ($\chi^2(1)$=7.355, p=0.02529). Because the interactions between the two features were so significant, results in Table 4.1 only include the Fixed Effects from the full model.

# Chapter 5

# DISCUSSION & FUTURE WORK

The goal of this study was to reproduce findings of previous studies in a standardized framework. This goal was accomplished. Per the summary of findings in (Gálvez, Gravano 2020), entrainment on pitch tended to have a negative impact on the trustworthiness of an avatar, whereas entrainment on volume tended to have a positive impact on the trustworthiness of an avatar. Note that configurations listed in Table 4.1 above are comparisons to the absence of an entrainment configuration, which most closely resembles the *against-static* format of (Gálvez, Gravano 2020). While the negative impact of pitch entrainment on trustworthiness was confirmed by this study's findings, the positive impact of volume entrainment on trustworthiness was contradicted by this study's findings, in Configuration 3 of Table 4.1. The interdependence of entrainment features was also confirmed by this study's results. Given such significant interdependence, one might expect that if entrainment on volume alone truly lowered trustworthiness, then it would be clear when comparing the volume-only model with the intercept-only model; however, such comparisons undermine the significance of Configuration 3 of Table 4.1 (much like the other volume-only configuration, which was discarded due to its insignificance even in the full model). While (Gálvez, Gravano 2020) did not find an effect of pitch disentrainment on trustworthiness, the results of this study showed that the effect was *also* negative; this lends additional evidence to the idea that observing an effect of entrainment on trustworthiness in one direction does not necessarily imply an effect in the opposite direction if the system instead disentrains on that same feature.

The layout of Table 4.1 is such that it is plain to see that the least trustworthy configuration was found to be when the agent disentrained on pitch, while the most trustworthy configuration was found to be when the agent entrained on both pitch and volume. This

offers a clear result to users, which can ideally be used to inform the entrainment configurations chosen for the spoken dialogue system they would like to train. Ideally, in future work, these analysis steps could be more fully automated, so that the R commands are automatically run, and follow an algorithm where the assumption of interdependence between features is first confirmed, before summarizing the full model; care should also be taken, however, to also discard data points which might seem misleadingly significant in the full model, by comparing feature-only models with the intercept-only as well. Automating such analysis is by no means a light task, though it is certainly a feasible and useful one. The other aspect that could use additional automation in future work is that of task/job management on the Appen platform. Jobs were manually created in the site's UI, though an API is available. Ideally, the Perception Evaluation Framework software suite would contain code to manage jobs on the platform that calls Appen's API, similar to the code it contains to manage jobs on mTurk. Management of the entrainment configurations on the SQS Queue is also a good target area for future automation. Once there are no longer active tasks running, the application can check the queue to see whether there are any configurations remaining on it. If there are, the application can generate new tasks for those configurations, and wait for those tasks to complete. Otherwise, it can proceed to the next step.

It is worth noting that asking users to evaluate the synthesized responses to their own recordings comes with its own set of trade-offs. One could conceive of a system where a single recording is used to seed multiple synthesized responses, each with a different entrainment configuration. Then, the same recording-to-response pair could be evaluated by a set of *different* participants so that the participant who uploaded the audio is excluded from the set of participants who evaluate the synthesized response. While this introduces numerous complications, there are strategies to overcome such complications, though they expand the scope of this work beyond what can be reasonably completed in the course of this paper.

Another area of future improvement is the qualification protocol. One could imagine future improvements where the app is configurable in the sense of being able to include or exclude certain demographics. If, for example, you are tailoring a spoken dialogue agent for a

specific audience, you might want to customize its entrainment configurations based on how that specific audience will perceive it. One could also imagine a scenario where demographics are used as fixed effects, such that even if including broad demographics, regression analyses would uncover disparities in how an entrainment configuration is perceived from one demographic to another. Additionally, being that alphanumeric identifiers separate from Appen's worker IDs are generated to disassociate workers from the data they create, the functionality for mTurk should be brought up to parity; currently, the mTurk interface logs worker IDs, the same way that the Speak Tool off of which it was built does.

The impact of stimulus as a random effect is another area to consider for future improvement. Being that each participant generates a unique audio file from which pitch or volume contours are extracted, such contours are unique per item, though could resemble another set of contours significantly or not at all. It's therefore possible that this either justifies also including a per-item random effect in the linear mixed models during analysis, or a more sophisticated mechanism that further treats two sets of contours in such a random effect as identical if they share a certain amount of information in common. This approach could also be used to differentiate entrainment and disentrainment entries as numerical, rather than categorical, based on the *degree* of entrainment or disentrainment; the complication with this latter approach being the reaffirmed finding that a feature's effect is not necessarily opposite just because it's entrained in the opposite direction, and so, treating entrainment and disentrainment of pitch as linear is likely not a valid solution, implying a different approach would be necessary to encompass numerical values.

The software framework has also not yet had duration entrainment incorporated into it. A future iteration will ideally include this, as well as options for all three entrainment *features* to utilize all three entrainment *types*. Enabling the option for all features and types will require groundwork on validating the basic *naturalness* bounds to which dialogue agent responses are constrained. For example, a scenario where a dialogue agent is disentraining on duration and volume, while entraining on pitch, might lead to highly unnatural outputs without proper restrictions in place around the limits of entrainment the dialogue agent is allowed to

perform. Such bounds are currently in place for pitch, but not volume. Implementing them for both volume and duration are important next steps. Additionally, the procedure used to synthesize volume convergence diverged from the approach used in (Levitan & Hirschberg, 2011); a future iteration might be to implement the synthesis step of volume convergence more closely to the one used for measurement *at the turn-level* in that paper, rather than adapting the one used at the session-level.

Lastly, future work could also focus on how findings differ by context. What might be a desirable vocal profile for a trustworthy dialogue agent in a sci-fi thriller video game is almost certainly not a desirable vocal profile for a trustworthy dialogue agent on your chosen banking website. The Percept-Eval framework could be made configurable such that researchers can incorporate the context in which their dialogue agent will operate. It can perhaps be contrasted with another context, so that both can be treated as fixed effects, for a richer analysis that verifies the output optimal entrainment configuration is indeed effective in the context in which it will be deployed.

# Chapter 6
# ETHICAL CONSIDERATIONS

When it comes to ethical considerations, the most concerning are those surrounding compensation through crowdsourcing platforms. Such platforms can be, and are, conceivably used to pay workers next to nothing for their time. The data collection component of the Perception Evaluation Framework being dependent on such platforms, it is not by accident that the data collection phase for the proof-of-concept entailed in this paper was conducted in the United States, where there is at least a minimum wage guaranteed to workers. Even so, exploitatively low wages are quite common on these platforms, even in locations where minimum wage is legally guaranteed, as has been well documented and explored for over a decade (Bederson & Quinn, 2011) (Fort & Cohen, 2011) (Snyder, 2010).

Moreover, mechanisms for unionization, and collective bargaining are virtually non-existent on these platforms. Combined with complaint mechanisms that are typically automated, this leaves workers even more powerless in scenarios where their work was undercompensated (for example, if a task took significantly longer than they were led to believe when completing it). Make no mistake, the absence of such mechanisms is not due to neglect, but by design; companies that crowdsource labor do actively lobby to prevent regulations that would otherwise require crowdsourced workers be protected by the same labor policies that guarantee the right to unionize and collectively bargain to employees (Chan, 2022). Other challenges which must be overcome to implement such regulations include difficulty identifying who the employer is, depending on the circumstances, as well as conflict with antitrust law (Johnston & Land-Kazlauskas, 2018). Absent such regulations, these issues seem to be endemic to *crowdsourcing* in general as a business model.

# Chapter 7

# CONCLUSION

This study introduced the Perception Evaluation Framework, which was successfully utilized, as a proof-of-concept, to reproduce results of a past study that correlated acoustic-prosodic features of speech with perceptual facets. The complex nature and multidimensionality of the interdependence between such acoustic-prosodic features renders the experimental design of such studies cumbersome, and the reproducibility of these experiments difficult. The aim of building this framework is to make such research more accessible and reproducible, so that researchers have the ability to more easily compare their findings to those of other researchers, in a consistent and clear manner. By replicating the results of a past study, it appears that the framework contains a sufficient amount of functionality to conduct such research reliably.

# BIBLIOGRAPHY

Appen Help Center. (n.d.). *How to create test questions.* https://success.appen.com/hc/en-us/articles/202702985-How-to-Create-Test-Questions. (Accessed: 2023-07-06)

Armitage, P., & Berry, G. (1994). *Statistical methods in medical research* (3rd ed.). Blackwell.

awsdocs. (2021). *aws-doc-sdk-examples.* https://github.com/awsdocs/aws-doc-sdk-examples. (Accessed: 2023-07-06)

Bederson, B. B., & Quinn, A. J. (2011). Web workers unite! addressing challenges of online laborers. In *Chi '11 extended abstracts on human factors in computing systems* (pp. 97–106). Association for Computing Machinery. doi: 10.1145/1979742.1979606

Benus, S., Trnka, M., Kuric, E., Martak, L., Gravano, A., Hirschberg, J., & Levitan, R. (2018). Prosodic entrainment and trust in human-computer interaction. In *Speech prosody 2018.*

Chan, W. (2022, Mar 11). Uber funds new lobbying group to deny rights for gig workers. *The Guardian*. Retrieved from https://www.theguardian.com/business/2022/mar/11/uber-funds-new-lobbying-group-to-deny-rights-for-gig-workers

Fort, K., Adda, G., & Cohen, K. B. (2011). Last words: Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, *37*(2), 413–420. doi: 10.1162/COLI_a_00057

Gálvez, R., Gauder, L., Luque, J., & Gravano, A. (2020). A unifying framework for modeling acoustic/prosodic entrainment: definition and evaluation on two large corpora. In *Sigdial.*

Gálvez, R., Gravano, A., Levitan, R., Trnka, M., & Hirschberg, J. (2020). An empirical

study of the effect of acoustic-prosodic entrainment on the perceived trustworthiness of conversational avatars. *Speech Commun.*, *124*, 46–67.

JeremyCCHsu. (2023). *Python-wrapper-for-world-vocoder.* `https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder/tree/master`. (Accessed: 2023-07-06)

Johnston, H., & Land-Kazlauskas, C. (2018). *Organizing on-demand: representation, voice, and collective bargaining in the gig economy* (ILO Working Paper No. 994981993502676). Geneva, Switzerland: International Labour Organization.

Levitan, R., Benus, S., Gálvez, R., Gravano, A., Savoretti, F., Trnka, M., ... Hirschberg, J. (2016). Implementing acoustic-prosodic entrainment in a conversational avatar. In *Interspeech.*

Levitan, R., Gravano, A., Willson, L., Benus, S., Hirschberg, J., & Nenkova, A. (2012). Acoustic-prosodic entrainment and social behavior. In *Naacl.*

Levitan, R., & Hirschberg, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Interspeech.*

Lubold, N., Pon-Barry, H., & Walker, E. (2015). Naturalness and rapport in a pitch adaptive learning companion. In *2015 ieee workshop on automatic speech recognition and understanding (asru)* (pp. 103–110).

Manson, J., Bryant, G., Gervais, M., & Kline, M. (2013). Convergence of speech rate in conversation predicts cooperation. *Evolution and Human Behavior*, *34*, 419–426.

Morise, M. (2016). D4c, a band-aperiodicity estimator for high-quality speech synthesis. *Speech Communication*, *84*, 57-65. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0167639316300413`

Morise, M. (2017). Harvest: A high-performance fundamental frequency estimator from speech signals. In *Proceedings of the interspeech 2017* (p. 2321-2325). Retrieved from `http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0068.html`

Morise, M., Yokomori, F., & Ozawa, K. (2016). World: a vocoder-based high-quality speech

synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, *E99-D*(7), 1877-1884. Retrieved from https://www.jstage.jst.go.jp/article/transinf/E99.D/7/E99.D_2015EDP7457/_article

Pérez, J., Gálvez, R., & Gravano, A. (2016). Disentrainment may be a positive thing: A novel measure of unsigned acoustic-prosodic synchrony, and its relation to speaker engagement. In *Interspeech.*

R Core Team. (2019). *car: Companion to applied regression.* https://cran.r-project.org/web/packages/car/index.html. (Accessed: 2023-07-06)

Snyder, J. (2010). Exploitation and sweatshop labor: Perspectives and issues. *Business Ethics Quarterly*, *20*(2), 187–213. Retrieved from http://www.jstor.org/stable/25702393

Song, C., Harwath, D., Alhanai, T., & Glass, J. (2022). Speak: A toolkit using amazon mechanical turk to collect and validate speech audio recordings. In *Proceedings of the language resources and evaluation conference* (p. 7253-7258). European Language Resources Association. Retrieved from https://aclanthology.org/2022.lrec-1.787

ST-AEDS-20180100_1, Free ST American English Corpus. (n.d.). *Available online.* (https://www.openslr.org/45/)

Thomason, J., Nguyen, H., & Litman, D. (2013). Prosodic entrainment and tutoring dialogue success. In *Aied.*

Wang, S., & Levow, G. (2011). Contrasting multi-lingual prosodic cues to predict verbal feedback for rapport. In *Acl.*

Yokozuka, T., Miyamoto, H., Kasai, M., Miyake, Y., & Nozawa, T. (2021). The relationship between turn-taking, vocal pitch synchrony, and rapport in creative problem-solving communication. *Speech Communication*, *129*, 33-40.

# Appendix A

# DATA TABLES

Table A.1: Data kept for analysis

| dp | subj | gen | ptch | vol | trst | disc | rsn | cond |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | M | 0 | 0 | 7 | no | n/a | none |
| 2 | 5 | M | 0 | 0 | 7 | no | n/a | none |
| 3 | 8 | M | 0 | 0 | 5 | no | n/a | none |
| 4 | 2 | M | 1 | -1 | 6 | no | n/a | e-ptch, d-vol |
| 5 | 8 | M | 1 | -1 | 5 | no | n/a | e-ptch, d-vol |
| 6 | 10 | M | 1 | -1 | 5 | no | n/a | e-ptch, d-vol |
| 7 | 3 | M | 1 | 0 | 4 | no | n/a | e-ptch |
| 8 | 8 | M | 1 | 0 | 4 | no | n/a | e-ptch |
| 9 | 8 | M | 1 | 0 | 3 | no | n/a | e-ptch |
| 10 | 7 | M | 1 | 1 | 6 | no | n/a | e-ptch, e-vol |
| 11 | 7 | M | 1 | 1 | 7 | no | n/a | e-ptch, e-vol |
| 12 | 9 | F | 1 | 1 | 6 | no | n/a | e-ptch, e-vol |
| 13 | 4 | M | 0 | 1 | 5 | no | n/a | e-vol |
| 14 | 10 | M | 0 | 1 | 6 | no | n/a | e-vol |
| 15 | 11 | M | 0 | 1 | 5 | no | n/a | e-vol |
| 16 | 3 | M | 0 | -1 | 5 | no | n/a | d-vol |
| 17 | 3 | M | 0 | -1 | 7 | no | n/a | d-vol |
| 18 | 8 | M | 0 | -1 | 4 | no | n/a | d-vol |
| 19 | 8 | M | -1 | 0 | 5 | no | n/a | d-ptch |
| 20 | 11 | M | -1 | 0 | 1 | no | n/a | d-ptch |
| 21 | 11 | M | -1 | 0 | 3 | no | n/a | d-ptch |
| 22 | 3 | M | -1 | -1 | 4 | no | n/a | d-ptch, d-vol |
| 23 | 6 | M | -1 | -1 | 6 | no | n/a | d-ptch, d-vol |
| 24 | 7 | M | -1 | -1 | 6 | no | n/a | d-ptch, d-vol |
| 25 | 6 | M | -1 | 1 | 5 | no | n/a | d-ptch, e-vol |
| 26 | 6 | M | -1 | 1 | 6 | no | n/a | d-ptch, e-vol |
| 27 | 7 | M | -1 | 1 | 5 | no | n/a | d-ptch, e-vol |

Table A.2: Data discarded before analysis

| dp | subj | gen | ptch | vol | trst | disc | rsn | cond |
|---|---|---|---|---|---|---|---|---|
| 1 | - | M | 0 | 0 | 7 | Yes | Audio Quality Issues | none |
| 2 | - | M | 0 | 0 | 5 | Yes | Audio Quality Issues | none |
| 3 | - | M | 0 | 0 | 4 | Yes | Background Noise | none |
| 4 | - | M | 0 | 0 | 6 | Yes | Audio Quality Issues; Background Noise | none |
| 5 | - | M | 1 | 0 | 6 | Yes | Audio Quality Issues | e-ptch |
| 6 | - | M | 1 | 0 | 7 | Yes | Audio Quality Issues; Correlation Issues | e-ptch |
| 7 | - | M | 1 | 0 | 5 | Yes | Audio Quality Issues | e-ptch |
| 8 | - | M | 1 | 0 | 7 | Yes | Audio Quality Issues; Correlation Issues | e-ptch |
| 9 | - | M | 1 | 0 | 7 | Yes | Audio Quality Issues | e-ptch |
| 10 | - | M | 1 | 0 | 7 | Yes | Silence | e-ptch |
| 11 | - | M | 1 | 0 | 7 | Yes | Dropout; Background Noise; Audio Quality Issues | e-ptch |
| 12 | - | M | 1 | 0 | 6 | Yes | Muffled Sound | e-ptch |
| 13 | - | M | 1 | 0 | 6 | Yes | Muffled Sound | e-ptch |
| 14 | - | M | 1 | 0 | 5 | Yes | Muffled Sound | e-ptch |
| 15 | - | M | -1 | 0 | 3 | Yes | Audio Quality Issues | d-ptch |
| 16 | - | M | -1 | 0 | 5 | Yes | Correlation Issues | d-ptch |
| 17 | - | M | -1 | 0 | 6 | Yes | Correlation Issues | d-ptch |
| 18 | - | M | -1 | 0 | 5 | Yes | Correlation Issues | d-ptch |
| 19 | - | M | -1 | 0 | 6 | Yes | Muffled Sound | d-ptch |
| 20 | - | M | -1 | 0 | 7 | Yes | Audio Quality Issues | d-ptch |
| 21 | - | M | -1 | 0 | 7 | Yes | Muffled Sound | d-ptch |
| 22 | - | M | 0 | 1 | 5 | Yes | Audio Quality Issues | e-vol |
| 23 | - | M | 0 | 1 | 7 | Yes | Muffled Sound; Background Noise | e-vol |
| 24 | - | M | 0 | -1 | 4 | Yes | Background Noise | d-vol |
| 25 | - | M | 0 | -1 | 7 | Yes | Dropout; Background Noise; Audio Quality Issues | d-vol |
| 26 | - | M | 1 | 1 | 7 | Yes | Correlation Issues | e-ptch, e-vol |
| 27 | - | M | 1 | 1 | 7 | Yes | Audio Quality Issues; Correlation Issues | e-ptch, e-vol |
| 28 | - | M | 1 | 1 | 5 | Yes | Audio Quality Issues | e-ptch, e-vol |
| 29 | - | M | 1 | 1 | 7 | Yes | Correlation Issues | e-ptch, e-vol |
| 30 | - | M | 1 | 1 | 6 | Yes | Muffled Sound | e-ptch, e-vol |
| 31 | - | M | -1 | -1 | 5 | Yes | Audio Quality Issues; Dropout | d-ptch, d-vol |
| 32 | - | M | -1 | -1 | 7 | Yes | Background Noise; Audio Quality Issues | d-ptch, d-vol |
| 33 | - | M | -1 | -1 | 7 | Yes | Dropout; Background Noise | d-ptch, d-vol |
| 34 | - | M | -1 | -1 | 6 | Yes | Dropout; Audio Quality Issues | d-ptch, d-vol |
| 35 | - | M | -1 | -1 | 4 | Yes | Muffled Sound | d-ptch, d-vol |
| 36 | - | M | -1 | -1 | 6 | Yes | Dropout; Background Noise | d-ptch, d-vol |
| 37 | - | M | 1 | -1 | 6 | Yes | Audio Quality Issues | e-ptch, d-vol |
| 38 | - | M | 1 | -1 | 5 | Yes | Dropout | e-ptch, d-vol |
| 39 | - | M | 1 | -1 | 7 | Yes | Audio Quality Issues | e-ptch, d-vol |
| 40 | - | M | 1 | -1 | 1 | Yes | Correlation Issues | e-ptch, d-vol |
| 41 | - | M | 1 | -1 | 7 | Yes | Dropout; Audio Quality Issues | e-ptch, d-vol |
| 42 | - | M | 1 | -1 | 6 | Yes | Background Noise | e-ptch, d-vol |
| 43 | - | M | 1 | -1 | 6 | Yes | Background Noise | e-ptch, d-vol |
| 44 | - | M | 1 | -1 | 5 | Yes | Muffled Sound | e-ptch, d-vol |
| 45 | - | M | 1 | -1 | 5 | Yes | Audio Quality Issues; Background Noise | e-ptch, d-vol |
| 46 | - | M | -1 | 1 | 5 | Yes | Dropout; Background Noise | d-ptch, e-vol |
| 47 | - | M | -1 | 1 | 7 | Yes | Muffled Sound | d-ptch, e-vol |
| 48 | - | M | -1 | 1 | 5 | Yes | Dropout; Background Noise | d-ptch, e-vol |
| 49 | - | M | -1 | 1 | 7 | Yes | Audio Quality Issues | d-ptch, e-vol |
| 50 | - | M | -1 | 1 | 5 | Yes | Muffled Sound | d-ptch, e-vol |