Summer 8-17-2017

Development and Validation for a Mobile Speech-in-Noise Audiometric Task

Tommy Peng
Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/eng_etds
Part of the Other Biomedical Engineering and Bioengineering Commons, and the Speech and Hearing Science Commons

Recommended Citation
https://openscholarship.wustl.edu/eng_etds/255

This Thesis is brought to you for free and open access by the Engineering and Applied Science at Washington University Open Scholarship. It has been accepted for inclusion in Engineering and Applied Science Theses & Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.
Development and Validation for a Mobile Speech-in-Noise Audiometric Task (Semantic Auditory Search) 

by

Tommy Qizhou Peng

A thesis presented to the School of Engineering of Washington University in St. Louis in partial fulfillment of the requirements for the degree of Master of Science

August 2017

St. Louis, Missouri
© 2017, Tommy Qizhou Peng
# Table of Contents

Table of Contents ......................................................................................................................... ii
List of Figures ................................................................................................................................. iv
Acknowledgments ............................................................................................................................ vi
ABSTRACT OF THE THESIS ......................................................................................................... viii

Chapter 1: Introduction ..................................................................................................................... 1

Chapter 2: Experimental Methods .................................................................................................. 5
  2.1 SAS Design and Implementation ............................................................................................... 5
    2.1.1 Preliminary Design ............................................................................................................. 5
    2.1.2 Audio Stimulus .................................................................................................................... 7
    2.1.3 Visual Stimulus ................................................................................................................... 10
    2.1.4 Questionnaire .................................................................................................................... 12
    2.1.5 Data Storage ....................................................................................................................... 13
  2.2 Presentation of SAS and QuickSIN to Human Participants .................................................... 14
    2.2.1 Presentation of SAS .......................................................................................................... 14
    2.2.2 Presentation of QuickSIN ................................................................................................ 17
  2.3 Prediction of QuickSIN Results Based on SAS Data ............................................................... 18
    2.3.1 Data Set ............................................................................................................................ 18
    2.3.2 Data Metrics ...................................................................................................................... 19
    2.3.3 Linear Regression Model .................................................................................................. 21
    2.3.4 Cronbach’s Alpha .............................................................................................................. 24

Chapter 3: Results ............................................................................................................................ 25
  3.1 Implementation of SAS ............................................................................................................ 25
  3.2 SNR Data ................................................................................................................................. 28
  3.3 Questionnaire Data .................................................................................................................. 31
  3.4 Linear Models of QuickSIN SNR50 and SAS Metrics ............................................................ 32
  3.5 Internal Consistency ................................................................................................................. 45

Chapter 4: Discussion ....................................................................................................................... 49
  4.1 Viability of Hearing Tests on Android Devices ........................................................................ 49
4.2 Effectiveness of Semantic Auditory Search ................................................................. 51
4.3 Future Directions ........................................................................................................... 54
References .......................................................................................................................... 56
Appendix A: Transcripts of Audio Tracks used in SAS ...................................................... 59
Appendix B: Target Word Lists used in SAS ..................................................................... 85
List of Figures

Figure 2.1 Preliminary design of SAS with multiple conversations ........................................ 6
Figure 2.2 A typical visual display of the SAS task during a task ........................................ 10
Figure 2.3 Questionnaire lists from all 5 tasks of SAS ..................................................... 12
Figure 2.4 A typical display of the questionnaire ............................................................... 13
Figure 2.5 A table of metrics measured during a task of SAS .......................................... 14
Figure 2.6 Summary of differences between SAS_v1 and SAS_v2 .................................. 16
Figure 2.7 Typical display of SAS_v2 during a task ......................................................... 17
Figure 2.8 Schematic for the creation of the correctness first response matrix ................... 19
Figure 2.9 The confusion matrix formed by checking Question Answer array of each subject against questionnaire answer key ................................................................. 21
Figure 3.1 Android amplitude presentation levels of attend and ignore track for participant 03 25
Figure 3.2 SNR calculated from recorded Android amplitudes compared to expected outcome 26
Figure 3.3 Reconstructed SNR compared to expected outcome ......................................... 27
Figure 3.4 Average percentage of last 20 responses being correct for different tasks of SAS .... 28
Figure 3.5 SNR at each target word for outliers, data from SAS_v1 ................................ 29
Figure 3.6 SNR at each target word for SAS_v1 subject data .......................................... 30
Figure 3.7 SNR at each target word for subject data from SAS_v2 .................................... 30
Figure 3.8 Confusion matrices for questionnaire answers of each task ............................. 31
Figure 3.9 Questionnaire statistics across all participants for all tasks of SAS_v1 ......... 31
Figure 3.10 Total number of correct first responses for each task as a linear function of QuickSIN SNR50; data from SAS_v1 ................................................................. 33
Figure 3.11 Linear model fit statistics for number of first response correct from different tasks; data from SAS_v1 ................................................................. 33
Figure 3.12 Comparing total number of correct first responses across all tasks for participants as a function of QuickSIN SNR50; data from SAS_v1 ........................................ 34
Figure 3.13 Linear model fit metrics for total number of correct first responses across all tasks as a function of QuickSIN SNR50; data from SAS_v1 ........................................ 34
Figure 3.14 Total number of correct first responses for each task as a linear function of QuickSIN SNR50; data from SAS_v1 ................................................................. 35
Figure 3.15 Linear model fit statistics for number of first response correct from different tasks, data from SAS_v2 ................................................................. 35
Figure 3.16 Comparing total number of correct first responses across all tasks for participants as a function of QuickSIN SNR50; data from SAS_v2 ........................................ 36
Figure 3.17 Linear model fit metrics for total number of correct first responses across all tasks as a function of QuickSIN SNR50; data from SAS_v2 ........................................ 36
Figure 3.18 Final SAS Score for each task as a linear function of QuickSIN SNR50, data from SAS ................................................................. 37
Figure 3.19 Linear model fit metrics for final SAS score; data from SAS_v1 ........................ 37
Figure 3.20 Comparing total final SAS scores across all tasks for participants as a function of QuickSIN SNR50; data from SAS_v1

Figure 3.21 Linear model fit metrics for total final SAS scores across all tasks as a function of QuickSIN SNR50; data from SAS_v1

Figure 3.22 Final SAS Score for each task as a linear function of QuickSIN SNR50; data from SAS_v2

Figure 3.23 Linear model fit metrics for final SAS score; data from SAS_v2

Figure 3.24 Comparing total final SAS scores across all tasks for participants as a function of QuickSIN SNR50; data from SAS_v2

Figure 3.25 Linear model fit metrics for total final SAS scores across all tasks as a function of QuickSIN SNR50, data from SAS_v2

Figure 3.26 Minimum SAS SNR for each task as a linear function of QuickSIN SNR50; data from SAS_v1

Figure 3.27 Linear model fit metrics for minimum SAS SNR, data from SAS

Figure 3.28 Comparing total minimum SAS SNR across all tasks for participants as a function of QuickSIN SNR50, data from SAS

Figure 3.29 Linear model fit metrics for total minimum SAS SNR across all tasks as a function of QuickSIN SNR50, data from SAS

Figure 3.30 Minimum SAS SNR for each task as a linear function of QuickSIN SNR50, data from SAS_v2

Figure 3.31 Linear model fit metrics for minimum SAS SNR, data from SAS_v2

Figure 3.32: Average SAS minimum SNR differences between normal and impaired

Figure 3.33 Comparing total minimum SAS SNR across all tasks for participants as a function of QuickSIN SNR50, data from SAS_v2

Figure 3.34 Linear model fit metrics for total minimum SAS SNR across all tasks as a function of QuickSIN SNR50, data from SAS_v2

Figure 3.35 Cronbach’s Alphas and their associated variances for SAS score and minimum SNR as the measurement metrics across both SAS_v1 and SAS_v2

Figure 3.36 Inter-task correlation for SAS Score; data from SAS_v1

Figure 3.37 Inter-task correlation for minimum SAS SNR performance; data from SAS_v1

Figure 3.38 Inter-task correlation for SAS Score; data from SAS_v2

Figure 3.39 Inter-task correlation for minimum SAS SNR performance; data from SAS_v2
Acknowledgments

I would like to express my sincere appreciation to my mentor, Dr. Dennis Barbour, for his vision, guidance and encouragement, without which this work would not have been possible. I would also like to thank my collaborator, Dr. Jonathan Peelle, for presenting interesting, yet practical challenges, which helped to refine this work to its current state. I am also grateful to my advisor, Dr. Barani Raman, for his intuition and support throughout my time at Washington University. I would like to acknowledge and thank David Song, Noah Ledbetter, Qihan Long and Chuck Holmes for contributing to the development of Semantic Auditory Search. I would like to thank all the members of the Laboratory of Neuroscience and Neuroengineering, all of whom I could always rely on for brainstorming. I would like to show my gratitude to Kristin Sobotka, who advised me to apply for both the Summer Undergraduate Research Award and the Summer Undergraduate Research Fellowship. I would also like to thank Glenn Reitz, Amanda Carr and the staff of the Department of Biomedical Engineering for their constant support and understanding.

Finally, I owe my deepest gratitude to my parents for their unwavering and unconditional support, which has made everything possible.

Tommy Peng

Washington University in St. Louis

August 2017
Dedicated to my parents.
ABSTRACT OF THE THESIS

Development and Validation for a Mobile Speech-in-Noise Audiometric Task

(Semantic Auditory Search)

by

Tommy Qizhou Peng

Master of Science in Biomedical Engineering

Washington University in St. Louis, 2017

Professor Dennis Barbour, Chair

Traditional speech-in-noise hearing tests are performed by clinicians with specialized equipment. Furthermore, these tasks often present contextually weak sentences in background babble, which are poor representations of real-world situations. This study proposes a mobile audiometric task, Semantic Auditory Search, which uses the Android platform to bypass the need for specialized equipment and presents multiple tasks of two competing real-world conversations to estimate the user’s speech-in-noise hearing ability. Through linear regression models built from data of seventy-nine subjects, three Semantic Auditory Search metrics have been shown to have statistically significant ($p < 0.05$) with medium effects sizes for predicting QuickSIN SNR50. The internal consistency of the task was also high, with a Cronbach’s alpha of 0.88 or more across multiple metrics. In conclusion, this preliminary study suggests that Semantic Auditory Search can accurately and reliably perform as an automated speech-in-noise hearing test. It also has tremendous potential for extension into automated tests of cognitive function, as well.
Chapter 1: Introduction

Hearing loss is one of the most common neurological disorders in America, affecting over 30 million Americans over the age of 12 (NIDCD 2016). While modern hearing aids and cochlear implants facilitate better sound reception under ideal conditions, the devices are not a complete solution in the real world due to the ubiquitous background noise and competing signals. In fact, suboptimal performance of hearing devices in noise is a major cause of hearing aid dissatisfaction (Kochkin 2011). Furthermore, the perception that devices do not work well in noise is one of the top barriers that negatively affect purchase decisions for hearing loss patients (Kochkin 2007).

While traditional pure-tone audiometry can offer estimates of hearing ability in quiet environments, older adults who have normal audiometric thresholds can have poor speech-in-noise perception (Gordon-Salant and Fitzgibbons 1993, Cruickshanks, Wiley et al. 1998). Furthermore, traditional audiograms often fail to accurately predict speech recognition performance in background noise (Souza, Boike et al. 2007, Anderson, Parbery-Clark et al. 2011, Souza, Arehart et al. 2011). In order to address these concerns, clinicians use speech-in-noise audiometric tests to quantitatively assess hearing ability in noise.

In general, speech-in-noise tests measure hearing ability in noise by simultaneously presenting a signal source of speech and a competing source of noise. The primary metric measured by speech-in-noise tests is the signal-to-noise ratio (SNR), which is the ratio between the intensity of the speech and the intensity of the noise. Commonly used clinical speech-in-noise tests such as QuickSIN and BKB-SIN use phonemically balanced sentences in increasingly higher-level background babble to estimate the SNR at which a person can report 50% of the key words in a
sentence (SNR50). Unfortunately, due to the phonemically restrictive nature of the presented signal, the sentences only contain limited contextual cues (McArdle and Wilson 2006) and thus differ significantly from the contextually rich conversations in noisy environments that the hearing-loss patients typically take part in. Therefore, SNR50 measurements from current clinical speech-in-noise tests offer only partial insight into the ability to listen to daily conversations in background noise.

Hearing impaired listeners have been shown to consistently perform worse than normal hearing listeners during speech-in-noise tests (Needleman and Crandell 1995). This observation is not surprising given that hearing impaired listeners typically have poor speech recognition ability in noise. Furthermore, this observation suggests that speech-in-noise tests can be used to screen patients for hearing loss. However, current clinical speech-in-noise tasks require delivery through an audiometer and the attention of a trained clinician. Such rigorous requirements on specialist attention and professional equipment not only increase the workload for clinicians but also significantly lower the accessibility of the tasks to the general population. Furthermore, due to the time commitment required from clinicians, many speech-in-noise tasks are designed to be short out of necessity, which limits the information that can be measured. These barriers also exist for many other psychometric assessments.

In order to address this general accessibility problem and simultaneously lower the workload for clinicians, investigators in recent years have begun investigating the viability of presenting automated psychometric tasks through a mobile platform. In particular, tests such as n-back, digit span and digit-symbol substitution have been successfully implemented on a mobile platform (DigitalArtefacts 2016). This transition towards mobile platform presentation is also particularly applicable to speech-in-noise tasks. Indeed, many of the current clinical audiometric tests, such
as HINT, were developed during a time when the general public had poorer access to audio playback interfaces. Due to the increasing popularity and the decreasing costs of mobile phones and tablets, which include audio playback capabilities, newly developed mobile tests are anticipated to be far more accessible to the general public. The computerized automation of data collection and data storage also allow for previously hard-to-obtain metrics, such as response reaction times for speech-in-noise tasks, to become efficiently captured. These advances offer opportunities for new testing paradigms to be designed and implemented. Indeed, researchers have used the iPad platform to administer novel psychometric tests to evaluate cognitive dysfunction (Marsha R. Zhang 2013).

In this thesis, a mobile platform was evaluated as a potential tool to present contextually rich speech-in-noise tests. To do this, (1) a novel automated speech-in-noise task, named Semantic Auditory Search (SAS), with adaptive presentation levels between speech and noise was designed and implemented, (2) SAS was presented to listeners through a tablet computer device and data were recorded, and (3) a quantitative model generated from the listener-generated data from SAS and the listener QuickSIN SNR50 scores was evaluated. The first objective aims to propose a contextually poised, real-world-like speech-in-noise test based on current speech-in-noise testing methods. The second objective aims to identify the difficulties and strengths of presenting speech-in-noise tests on the mobile platform. The final objective aims to examine the reliability of the newly proposed test.

However, due to the diversity in hardware and software across mobile platforms, calibration and reliability become central problems that need to be resolved in future studies. Calibration for traditional speech-in-noise tasks generally relies on the fact that the listeners are in an audiometric booth or room. This is not the case for SAS, which can be presented under a variety
of real-world background noise due to its mobility. Thus, a new calibration regime must be implemented to ensure that the generated data accurately reflect the underlying patient capabilities. Furthermore, a more rigorous test-retest reliability metric must also be established for SAS. Due to the contextually rich nature of the sound sources in the task, listeners may learn significant portions of the task and thus act based on the gathered information if presented with the same stimulus conditions. Thus, the definition of a SAS re-test cannot be a repeat of presentation and must be investigated further.
Chapter 2: Experimental Methods

This chapter describes the SAS task development process, the subsequent experimentation methods, and the data analysis approaches used in this thesis. The rationale of the design decision or methods will be explored during the results section of the thesis.

2.1 SAS Design and Implementation

This segment of the thesis contains an overview of the SAS task and its implementation.

2.1.1 Preliminary Design

Search tasks are well known in vision science (Treisman and Gelade 1980, Wolfe 1994, Wolfe, Alvarez et al. 2011) and have been used in visual testing tasks, such as the Useful Field of View, for measuring deficits in the central visual field (Ball, Beard et al. 1988). However, search tasks have been less comprehensively studied in auditory science, with the majority of studies focusing on contrasts between simple acoustic features (Cusack and Carlyon 2003, Lallier, Donnadieu et al. 2013) or interactions between attentional and semantic effects (Treisman and Squire 1974). Semantic auditory search is designed to bridge this gap as a complex auditory search task designed to identify subtle deficits in both auditory and cognitive processing.

In the original formulation of the task, target cues are presented to users in one of several conversational streams within a variable time interval. Each conversation is represented on screen with a black dot. The user’s auditory search space is presented as a white circular area. The users are able to alter the spatial positions of the conversations in this “cocktail party” by rotating their auditory search space to simulate a head turn. Target cues are displayed in a button or buttons on the top of the screen, and are typically spoken/orthographic words, phrases, concepts, questions, rhymes, etc. The user is instructed to first press the button when the word related to the cue appears in one of the conversations, then drag and drop the button on top of the
black dot representation of the conversation from which the word related to the cue originated.

The result is a task with liberal user-directed control, an important factor for maintaining motivation in users (Garris, Ahlers et al. 2002), while also maintaining multiple strategies for the test administrator to modulate task difficulty based on user performance.

---

**Figure 2.1** Preliminary design of SAS with multiple conversations. The auditory search space is the large white circle in the background. The black dots are the visual representations of the locations of the presented conversations relative to the user. The direction of the user within the auditory search is indicated by the pointed “nose” structure on the user sprite. As shown in the figure, one source is on the user’s right and one source is in front of the user, slightly to the right. The target word “cut” is a word that will be presented by one of the two conversations. The user is asked to listen for the word and drag the button to the dot of the sound source that the word came from.
This original version of SAS, with large number of task variations and in-task variables, was designed for evaluating audiometric testing effectiveness over a large number of subjects. For the purposes of this study, SAS was modified to a two-conversation, non-spatial, auditory search task with fewer in-task variations, therefore making it more suitable for a smaller number of subjects. In this reduced version, the cues presented on the button were simplified from semantic cues to simple orthographic cues (e.g., “dog” for dog in the reduced version instead of “canine” for dog).

2.1.2 Audio Stimulus
Ten radio program recordings were chosen from National Public Radio broadcasts along with respective transcripts, shown in Appendix A. The individual recordings were five minutes in length. The recordings are stored as WAV files with sampling rates of 16000 samples/s. The root mean square amplitudes of all recordings were equalized to an average of 0.7. The recordings were then placed into five distinct pairs, with each recording used exactly once. One recording from each pair was designated to be the signal audio stream or “attend track”, and the other recording from the pair was designated to be the noise audio stream or “ignore track”. A pair of recordings was presented continuously and simultaneously through both the left and right audio channels, and therefore to both ears, during iterations within the SAS task. Whether the stimuli delivered were diotic or dichotic depended on the experimental design. SAS as implemented contains a total of five iterations, also designated as “tasks” or “task levels,” each corresponding to one pair of recordings. The presentation order of the iterations was the same for each test subject.

An array of the start and end times of individual words in milliseconds relative to the start of the recording was generated from the transcript and the audio file of each recording. From this array,
a selection of contextually important or phonologically interesting words, also known as a “word list,” was selected for each recording, with between 2 to 8 seconds in between the onset of each successive word, also known as “target word”. The word lists used in SAS can be found in Appendix B. The word list of the attend track of each task was used to generate a set of successive prompts to which the listeners responded. The timing of the response from the listener can then be judged relative to a correct timing window. The correct timing window for SAS was set to be between 0.5 seconds before and 2 seconds after the start of target word presentation as indicated by the wordlist. Responses to target words before correct timing window were considered too early and responses within the timing window were considered on time. Responses prior to the start of the target word were allowed because contextual information makes it possible to anticipate the target word, which could result in a correct response prior to actual word delivery. In order to incentivize the listener to listen at the lowest SNR, a simple scoring mechanism was implemented as discussed in section 2.1.2. If the response was too early, the target word did not change, but a score penalty was incurred. If the response occurred within the timing window, the target word was updated to the next target word and the listener’s score increased. The task automatically updated the target word to the next target word if no response occurred during the timing window.

SNR was used to quantify the auditory environment during word presentation. SNR for a particular auditory environment, measured in dB, can be calculated using the amplitude of the signal $A_{signal}$ and the amplitude of the noise $A_{noise}$ (2.1). For Android devices on Android API level 19, amplitude values for native audio playback classes were directly proportional to the amplitudes of the output sound wave. This was empirically verified by presenting a 1 kHz tone at different amplitudes using the Android native AudioTrack class through an ASUS Nexus 7.
device connected to an oscilloscope. This relationship between amplitude of sound wave $A_{\text{sound}}$ and Android amplitude of the sound $V_{\text{sound}}$ can be summarized in (2.2), which can then be used to derive (2.3).

$$SNR_{dB} = 20 \log_{10} \left( \frac{A_{\text{signal}}}{A_{\text{noise}}} \right)$$

(2.1)

$$A_{\text{sound}} = k V_{\text{sound}}$$

(2.2)

$$SNR_{dB} = 20 \log_{10} \left( \frac{k V_{\text{signal}}}{k V_{\text{noise}}} \right) = 20 \log_{10} \left( \frac{V_{\text{signal}}}{V_{\text{noise}}} \right)$$

(2.3)

The initial Android amplitude of the attend track was set to 0.65 while the initial Android amplitude of the ignore track was set to 0.35. This arrangement created a relatively large initial SNR of 5.38 dB to assist the listener in identifying the attend track at the beginning of each task. The SNR of the task adapts based on the correctness of the responses given by the listener according to a one-up/two-down paradigm, which increases the SNR if the listener responds incorrectly and decreases the SNR if the listener responds correctly twice in a row. The magnitude of the increase and decrease in SNR, also known as step size, was set to be 1.25 dB. The total Android amplitude of the two tracks remained close to 1 throughout the task.
2.1.3 Visual Stimulus

![Image](image.png)

**Figure 2.2** A typical visual display of the SAS task during a task. The response button can be seen at the top with the target word “fiscal”. The Android amplitudes of the ignore and attend tracks are visually represented as two sliders, with the top of the slider representing an Android amplitude of 1, and the bottom of the slider representing an Android amplitude of 0. The onscreen orthographic word representation can be seen in cyan. The current SNR of the task, which is approximately 5.38 dB, can be seen on the left, between the Android amplitude indicators and the target button. In the current implementation of SAS, the sliders cannot be changed manually.

The individual target word prompts from the word list of a particular task level were visually displayed on a button, also known as the “response button.” Once the listener starts a task, the first target word on the word list for the task is displayed on the response button. The target word on the response button updates when the button is pressed during the correct timing window of the target word or when the in-task time exceeds the correct timing window of the target word. For each task, a hyphen (“-”) was displayed on the response button following the timing window for the last target word to indicate to the listener that the task was ending.
The task interface provides two real-time sliders, which are visual representations of the Android amplitudes of the attend and ignore tracks. For both tracks, the bottom of the slider indicates an Android amplitude level of 0, and the top of the slider indicates an Android amplitude level of 1. In this implementation of SAS, the sliders are only controlled by task performance and are not user-adjustable.

The task also included a user feedback text displayed above the response button based on the correctness of a listener response. “Nice! Your response was on-time” was displayed in green in the case in which the listener responded during the correct timing window. “Your response was too early” was displayed in red when the listener responded before the correct timing window. “No on-time response detected, word refreshed” was displayed in cyan when the listener failed to respond within the correct timing window.

A basic score was implemented to incentivize the listener to perform the task at a lower SNR. Each presented word has a score value that is determined by a combination of the Android amplitude of the attend track \( V_{\text{attend}} \) and the SNR of the two tracks \( SNR_{dB} \) at the time of word presentation.

\[
\text{current score} = (600 + 100 \times V_{\text{attend}}) - (SNR_{dB} \times 50) \quad (2.4)
\]

A total score metric was then calculated from the current score. If the listener correctly responded to the target word, the current score was added to the total score of the task. If the listener was too early during the response, 25% of the current score was deducted from the total score. Both the current and total scores were displayed continuously on screen during the task.
2.1.4 Questionnaire
A short 20-word questionnaire pertaining to the content of the stories was presented after each task level. The words were contextually significant words chosen from three groups: 7 words from the word list of the attend track, 8 words from the ignore track, 5 words from the attend track but were not in the word list of the attend track. The ordering of the words was randomized for each list. The words within the lists were delivered in the same order to every subject.

<table>
<thead>
<tr>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
<th>Task 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>revenue</td>
<td>exchange</td>
<td>fluent</td>
<td>scary</td>
<td>sketch</td>
</tr>
<tr>
<td>meeting</td>
<td>allocate</td>
<td>investor</td>
<td>torches</td>
<td>linked</td>
</tr>
<tr>
<td>abuse</td>
<td>wave</td>
<td>obsessed</td>
<td>libbaster</td>
<td>approval</td>
</tr>
<tr>
<td>capital</td>
<td>promote</td>
<td>episodes</td>
<td>interactive</td>
<td>scholars</td>
</tr>
<tr>
<td>stalled</td>
<td>minister</td>
<td>idea</td>
<td>defense</td>
<td>testament</td>
</tr>
<tr>
<td>immigration</td>
<td>sense</td>
<td>market</td>
<td>planes</td>
<td>detector</td>
</tr>
<tr>
<td>democratic</td>
<td>routinely</td>
<td>character</td>
<td>goggles</td>
<td>western</td>
</tr>
<tr>
<td>speaker</td>
<td>mandatory</td>
<td>lincoln</td>
<td>situation</td>
<td>telegraph</td>
</tr>
<tr>
<td>stained</td>
<td>counterpart</td>
<td>accurate</td>
<td>elevator</td>
<td>cognitive</td>
</tr>
<tr>
<td>budget</td>
<td>sunshine</td>
<td>hurricane</td>
<td>transparent</td>
<td>insane</td>
</tr>
<tr>
<td>morning</td>
<td>crisis</td>
<td>product</td>
<td>senator</td>
<td>popularity</td>
</tr>
<tr>
<td>proposal</td>
<td>building</td>
<td>hollywood</td>
<td>castle</td>
<td>fish</td>
</tr>
<tr>
<td>oval</td>
<td>money</td>
<td>mafia</td>
<td>characters</td>
<td>human</td>
</tr>
<tr>
<td>reported</td>
<td>together</td>
<td>economy</td>
<td>musician</td>
<td>absorb</td>
</tr>
<tr>
<td>flare</td>
<td>doctors</td>
<td>winning</td>
<td>drones</td>
<td>benefits</td>
</tr>
<tr>
<td>leadership</td>
<td>slack</td>
<td>terms</td>
<td>lawmakers</td>
<td>headline</td>
</tr>
<tr>
<td>resolution</td>
<td>orders</td>
<td>national</td>
<td>casualties</td>
<td>conflict</td>
</tr>
<tr>
<td>shutdown</td>
<td>psychiatric</td>
<td>soup</td>
<td>senior</td>
<td>culture</td>
</tr>
<tr>
<td>entitlement</td>
<td>talks</td>
<td>comparatively</td>
<td>congress</td>
<td>author</td>
</tr>
<tr>
<td>homes</td>
<td>choreography</td>
<td>measure</td>
<td>legitimate</td>
<td>civilization</td>
</tr>
</tbody>
</table>

Figure 2.3 Questionnaire lists from all 5 tasks of SAS. Green words (e.g., Revenue) are from the word list of the attend track, red words (e.g., Capital) are from the ignore track, and blue words (e.g., Meeting) are from the attend track but are not in the word list of the attend track.

The questionnaire presented individual words from questionnaire lists in the order found in Table 1. For each word, the questionnaire presented two alternatives: “yes” was to be selected if the word was from the attend track, “no” was to be selected if the word was not from the attend track. Each word of the questionnaire required an answer before the display of the “Next” button. The answer selection of a particular word could not be changed after “Next” was pressed for that word. Therefore, the questionnaire is a modified “yes-no” version of a two-alternative forced choice task.
Figure 2.4 A typical display of the questionnaire. The word in question (Revenue) is displayed at the top in cyan. There are two buttons: “yes” and “no.” After selection, the text on the button turns green, indicating that the choice has been selected. The user can then change his or her response if desired. Next is pressed once the selection is final, and saves the selection.

2.1.5 Data Storage
Data collected during each task of SAS was pushed to a Parse.com server after completion of the level. The data was stored in a NoSQL format on a password-protected Parse account. The data can be accessed and downloaded as JSON files, which can then be analyzed offline.

The following are a description of the metrics saved after every task of SAS. Many of these metrics were stored within arrays, where each entry into the array corresponds to the value of the metric for a press of the response button or answer to a word in the questionnaire. The entries were concatenated to the end of the array in chronological order.
<table>
<thead>
<tr>
<th>Name of Metric</th>
<th>Description of Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word Index (Array)</td>
<td>The index of the current target word in word list. Indexing starts at 0.</td>
</tr>
<tr>
<td>Amplitude of Ignore Track (Array)</td>
<td>The Android AudioTrack amplitude for the ignore track, with a minimum of 0 and a maximum of 1.</td>
</tr>
<tr>
<td>Amplitude of Attend Track (Array)</td>
<td>The Android AudioTrack amplitude for the attend track, with a minimum of 0 and a maximum of 1.</td>
</tr>
<tr>
<td>Response Delay (Array)</td>
<td>The time in milliseconds between the start of the word as given by the word list and the response of the listener.</td>
</tr>
<tr>
<td>Correctness (Array)</td>
<td>The correctness of response button pressing timing. -1 if the attempt is too early, 0 if it is within the correct timing window, 1 if no response was received within the correct timing window.</td>
</tr>
<tr>
<td>Question Answer (Array)</td>
<td>The answer for a particular question in the questionnaire. 0 if “yes” was selected, 1 if “no” was selected.</td>
</tr>
<tr>
<td>User</td>
<td>The unique string identifier for the current listener.</td>
</tr>
<tr>
<td>UpdatedAt</td>
<td>The date and time that the data were pushed to Parse.</td>
</tr>
<tr>
<td>Current Score (Array)</td>
<td>The score assigned by the current target word.</td>
</tr>
<tr>
<td>Total Score (Array)</td>
<td>The total score achieved at the time of each response.</td>
</tr>
</tbody>
</table>

**Figure 2.5** A table of metrics measured during a task of SAS.

### 2.2 Presentation of SAS and QuickSIN to Human Participants
This section describes the protocol and approaches used to obtain human participant data using the SAS and QuickSIN techniques.

#### 2.2.1 Presentation of SAS
A total of 79 participants were recruited to perform on two different versions of the SAS task.

In SAS_v1, 37 participants over the age of 18 with fluent English understanding were recruited from the Washington University in St Louis community. The task was conducted inside an Acoustic Systems RE series sound booth and on an ASUS Nexus 7 tablet device with Android.
API level 19. The Nexus 7 device sound output amplitude was set to be 4 out of 15. The task was presented through a set of Audio-Technica ATH-M50 circumaural headphones.

The participants were presented with 5 task levels of SAS and a questionnaire at the end of each task level. All stimuli for SAS_v1 were delivered diotically, with both attend and ignore streams presented to both ears. A set of instructions was given before the start of SAS.

_Imagine you are in a room where people are having two different conversations. You want to listen to one conversation, which is also called the “attend” conversation, and ignore the other conversation. The “attend” conversation that you want to listen to will be easy to hear at first, because it will be louder than the conversation that you want to ignore._

_There will be a response button near the top of the screen which will display a word that will come up in the “attend” conversation that you want to listen to. Once you hear the displayed word in the “attend” conversation, please hit the response button as soon as possible. You may respond to the word as many times as you like as long as it is still on the button._

_Your total score for any particular level will be displayed in the top right of the screen. Please attempt to achieve the highest total score during each level. The Final button on each level will be a hyphen followed by a blank._

_There will be 5 levels in total. There will be a test on the content of the “attend” conversation after each level. Feel free to take breaks in between levels._

The participants were informed of the two competing conversations, the need to focus on each attend conversation, and the potentially varied relative sound level of the two conversations.

Participants were informed that the attend conversation was relatively louder at first, but could change in relative loudness. Participants were instructed to note the target word, and to immediately press the response button after the presentation of the target word in the attend conversation. The scoring metric was also described. In particular, the participants were instructed to achieve the highest total score during each task.

Another set of instructions was given after the audio portion of SAS and before the questionnaire over the first task level. The participants were informed that 20 words would appear individually, that the words originated from either the attend or ignore conversation, and that they were
instructed to press “yes” if the word originated from the attend conversation and “no” if the word originated from the ignore conversation. Participants were informed that the task would continue only if a selection were made.

You will now be presented with 20 words. The words will have come from the “attend” or “ignore” conversation. Please press “yes” if you think the word comes from any part of the attend stream, and press “no” if you think that the word did not come from the attend stream.

A further 42 subjects were given a modified 6 task version of SAS (SAS_v2). Four separate dichotic tasks, which delivered the attend conversation exclusively to one ear and ignore conversation exclusively to the other ear, were followed by 2 diotic tasks. The ordering of the tasks was: left attend, right attend, left attend, right attend, diotic, diotic. The two diotic tasks in SAS_v2 are similar to the first two tasks found in SAS_v1. As the participant group was expected to be older in SAS_v2, the user interface was also modified such that the target word button was enlarged, and the attend and ignore sliders were hidden. The questionnaire was not presented after each task in SAS_v2. The differences between the two tasks can be found in the following table.

<table>
<thead>
<tr>
<th>Task</th>
<th>Attend Track</th>
<th>Ignore Track</th>
<th>Attend Track Ear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>1</td>
<td>2</td>
<td>Diotic</td>
</tr>
<tr>
<td>Task 2</td>
<td>3</td>
<td>4</td>
<td>Diotic</td>
</tr>
<tr>
<td>Task 3</td>
<td>5</td>
<td>6</td>
<td>Diotic</td>
</tr>
<tr>
<td>Task 4</td>
<td>7</td>
<td>8</td>
<td>Diotic</td>
</tr>
<tr>
<td>Task 5</td>
<td>9</td>
<td>10</td>
<td>Diotic</td>
</tr>
<tr>
<td>Task 6</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
</tbody>
</table>

| Task 1 | 9 | 10 | Left |
| Task 2 | 11 | 12 | Right |
| Task 3 | 5 | 6 | Left |
| Task 4 | 7 | 8 | Right |
| Task 5 | 1 | 2 | Diotic |
| Task 6 | 3 | 4 | Diotic |

Figure 2.6 Summary of differences between SAS_v1 and SAS_v2.
2.2.2 Presentation of QuickSIN
For all SAS participants, QuickSIN was presented through a Toshiba Portege R700-S1310 laptop with Windows 7 operating system amplitude set to 14 and VLC audio player amplitude set to 100. The task was delivered through AudioTechnica ATH-M50 circumaural headphones. The participants were seated inside an Acoustic Systems RE series sound booth and told to listen and repeat the sentences to their best abilities in the task. The instructions given were similar to that found in the QuickSIN user manual (2006).
Imagine that you are at a party. There will be a woman talking and several other talkers in the background. The woman’s voice is easy to hear at first, because her voice is louder than the others. Repeat each sentence the woman says. The background talkers will gradually become louder, making it difficult to understand the woman’s voice, but please guess and repeat as much of each sentence as possible.

Two standard equivalent lists in QuickSIN, list 3 and 4, were used during data collection. All participants were presented with the audio tracks associated with the two lists, in the same order. A total of 6 sentences populate each list, with 5 key words in each sentence. The listener’s SNR50 metric, a SNR at which the listener can identify 50% of the key words in a sentence, was calculated by subtracting the number of key words the listeners identified correctly in the whole list from 25.5. The QuickSIN SNR50 metric was calculated as instructed in the QuickSIN user manual. For the two lists, according to the QuickSIN user manual, the averaged SNR50 is accurate to 1.9 dB, 1.6 dB and 1.3 dB at the 95%, 90% and 80% confidence levels, respectively.

For the SAS_v2 participant group, QuickSIN was delivered through an audiometer, calibrated as per the QuickSIN instruction manual. Similarly, two standard equivalent lists were used to obtain an average SNR50 for the participants.

2.3 Prediction of QuickSIN Results Based on SAS Data

2.3.1 Data Set
The original data was exported from the Parse.com server as a single JSON file. The individual task data were stored as individual arrays. Each of these arrays contains subarray of metrics discussed in section 2.1.5. For SAS_v1, there were a minimum of 5 arrays for each subject, and for SAS_v2 there were a minimum of 6 arrays for each subject, which corresponds to the number of tasks per SAS version. The JSON file was then parsed into MatLab using jsonLAB and manipulated into a cell array with the rows corresponding to each subject, and the columns corresponding to each task. Data from errored or crashed tasks were discarded during this
transform. The metrics discussed in the following sections can be calculated from the data found in this cell array.

**2.3.2 Data Metrics**

To investigate the response behavior of subjects throughout the tasks, a variety of *first response matrices* were computed for each task for the subject. The values found in the first response matrices corresponded to the value of the metric during the first response of the subject to target words found in that task. For example, the *correctness first response matrix* was populated by using the data from the Correctness and Word Index arrays of the task. A value of -1 was assigned to the case in which the subject’s response was too early, 0 if the subject was on-time, and 1 if the subject was too late.

| Word Index | 1 | 2 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 9 | 10 | ...
|------------|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Correctness| 1 | -1| 0 | 0 | 0 | 0 | 1 | 0 | 0 | -1| 1 | 0 | ...

**Figure 2.8** Schematic for the creation of the correctness first response matrix. The original data matrix (top) is transformed into the correctness first response matrix (bottom). Note how there are multiple responses for target words 2 and 9 (in red) originally, but the first response matrix only takes the metric (correctness) value during the very first response and drops the values in red.

Similarly, *SNR first response matrices* and *Response Delay first response matrices* are calculated from the calculated SNR at each world and the Response Delay array for the raw data respectively.

During analysis performed in section 3.1, it was noted that there were some discrepancies between the SNR calculated using equation 2.3, Amplitude of Ignore Track and Amplitude of Attend Track arrays, and the expected amplitude changes as per the one-up-two-down
psychoacoustic testing strategy. These discrepancies were caused by an error in the code for SAS. SAS was saving data before SNR update for cases when the user was too early or too late (Correctness array entry -1 or 1), but was saving data after SNR update for cases when the user was on time. This resulted in a shifted graph when comparing expected amplitude changes and the actual amplitude changes from SNR calculated from Amplitude of Ignore Track and Amplitude of Attend Track. Luckily, later in section 3.1, the on-up-two-down psychoacoustic strategy was found to be correctly implemented within the task through display of real-time SNR within the application. However, since the discrepancies were found after human subject data acquisition, the SNR array for each task had to be reconstructed from the correctness array and initial SNR conditions.

The SNR of the last k responses of the task were arithmetically averaged to form the \( k \)-average end SNR to quantify the listening environment presented to the subjects at the end of each task. The minimum SNR for each participant during each task was also found by finding the minimum of each SNR matrix. The response delay matrix was changed from units of milliseconds to units of seconds by dividing the individual raw values for each response by 1000.

The questionnaire lists were reduced to a numerical representation as an array of questionnaire answer keys. In this array, 1’s represented words which were both in the word list and presented in the attend track or were only presented in the attend track but not in the word list, and 0’s represented words which originated in the ignore track. The Question Answer arrays for each subject could then be checked against the questionnaire answer key to populate a confusion where the true values are the answer keys and the predictions are done by the subjects. The values found inside the confusion matrix were used to calculate the questionnaire true positive
rate (QTPR), questionnaire true negative rate (QTNR), questionnaire false positive rate (QFPR), and questionnaire false negative rate (QFNR).

<table>
<thead>
<tr>
<th></th>
<th>Question Answer Attend</th>
<th>Question Answer Ignore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answer Key Attend</td>
<td>True Positive</td>
<td>False Negative</td>
</tr>
<tr>
<td>Answer Key Ignore</td>
<td>False Positive</td>
<td>True Negative</td>
</tr>
</tbody>
</table>

**Figure 2.9** The confusion matrix formed by checking Question Answer array of each subject against questionnaire answer key.

\[
QTPR = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (2.5)
\]

\[
QTNR = \frac{\text{True Negative}}{\text{False Positive} + \text{True Negative}} \quad (2.6)
\]

\[
QFPR = \frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}} \quad (2.7)
\]

\[
QFNR = \frac{\text{False Negative}}{\text{True Positive} + \text{False Negative}} \quad (2.8)
\]

Finally, the accuracy of questionnaire response for all subjects can be calculated at each task.

\[
\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Negative} + \text{False Positive} + \text{True Negative}} \quad (2.9)
\]

### 2.3.3 Linear Regression Model

Regression models are often used to observe the relationship between a scalar dependent variable and one or more explanatory variables. Linear regression models are regression models built under the stipulation that the underlying function is linear. Linear regression can be generalized to the following model.

**QuickSIN SNR50 for subject i**

\[
\text{QuickSIN SNR50 for subject } i = \beta_0 + \beta_1 \text{SASMetric}(i, 1) + \beta_2 \text{SASMetric}(i, 2) + \cdots + \beta_p \text{SASMetric}(i, p)
\]
Where $i = 1, \ldots, n$ is the index for subjects, and there are $p$ number of explanatory variables

$$ y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} SNR50_1 \\ SNR50_2 \\ \vdots \\ SNR50_n \end{pmatrix} $$

$$ X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} = \begin{pmatrix} SASMetric_1^T \\ SASMetric_2^T \\ \vdots \\ SASMetric_n^T \end{pmatrix} = \begin{pmatrix} 1 & SASMetric_{11} & \cdots & SASMetric_{1p} \\ 1 & SASMetric_{21} & \cdots & SASMetric_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & SASMetric_{n1} & \cdots & SASMetric_{np} \end{pmatrix} $$

$$ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} $$

Ordinary least squares method estimates $\beta$ by:

$$ \hat{\beta} = (X^T X)^{-1} X^T y \quad (2.10) $$

For SAS related studies, the QuickSIN SNR50 of each participant is the scalar dependent variable, and the participant’s performance during SAS_v1 or SAS_v2 are the explanatory variables.

Traditionally, there are many measures of regression model performance. For linear regression models, the goodness of fit is often characterized by the R-squared value, which describes the proportion of variation in the dependent variable that is due to variation in the explanatory variables.

**Sum of Squares for Model (SSM)**

$$ Sum of Squares for Error (SSE) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2.11) $$

$$ Sum of Squares for Error (SSE) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2.12) $$

**Corrected Sum of Squares Total (SST)**

$$ Corrected Sum of Squares Total (SST) = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad (2.13) $$

Where $\hat{y}_i$ is the estimate of $y_i$ by the model and $\bar{y}$ is the mean of $y$. 

22
The F-test statistic is also an important metric to consider when evaluating the validity of the linear model. For linear regression models, the null hypothesis of the F-test is that all the values of $\beta$ are 0, corresponding to the case where none of the explanatory variables have an effect on the model, and the alternative hypothesis is that at least one explanatory variables has an effect on the model.

\[ \text{Null Hypothesis (} H_0 \text{)}: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \]

\[ \text{Alternative Hypothesis (} H_1 \text{)}: \beta_j \neq 0, \text{for at least one value of } j \]

The F-test statistic can then be calculated as the following.

\[ \text{Corrected Degrees of Freedom for Model (} DFM \text{)} = p - 1 \]  
(2.15)

\[ \text{Degrees of Freedom for Error (} DFE \text{)} = n - p \]  
(2.16)

\[ \text{Mean of Squares for Model (} MSM \text{)} = \frac{SSM}{DFM} \]  
(2.17)

\[ \text{Mean of Squares for Error (} MSE \text{)} = \frac{SSE}{DFE} \]  
(2.18)

\[ F \text{ statistic} = \frac{MSM}{MSE} \]  
(2.19)

Computationally, the models were solved using the fitlm function in the MATLAB Statistics and Machine Learning Toolbox using the method of least squares fit as described above. The output included R-squared, F-test statistic and p-value for F-test statistic for each linear model built.

However, precautions must be taken when making multiple statistical inferences simultaneously. In what is known as the multiple comparisons problem, the large number of statistical tests, performed simultaneously, can result in p-values less than threshold through chance alone. This effect can be somewhat mitigated through the Bonferroni correction.
\[ p_{task} = \frac{p_{total}}{\text{number of tasks}} \]  

(2.20)

For the statistical analysis performed in this study, the total p-value threshold is 0.05. Therefore, p-value for individual tasks after correction is 0.01 for the 5 SAS_v1 tasks, and 0.00833 for the 6 SAS_v2 tasks.

### 2.3.4 Cronbach’s Alpha

The Cronbach’s Alpha is a popular measure of internal consistency within behavioral science. In this case, internal consistency measures how consistently different tasks found within SAS measure a subject’s speech-in-noise hearing ability.

Given variables \( x_1, \ldots, x_k \), and \( x_0 = \sum_{j=0}^{k} x_k \)

\[
\alpha = \frac{k}{k-1} \left(1 - \frac{\sum_{j=1}^{k} \text{var}(x_j)}{\text{var}(x_0)}\right)
\]

(2.20)

For SAS, a select metric from each task, or \( x_k \) in the above formulae, can be consider an individual assessment of the subject’s speech-in-noise hearing ability, resulting in a \( k \) of 5 for SAS_v1 and a \( k \) of 6 for SAS_v2. The sum of that metric across all tasks is therefore \( x_0 \).
Chapter 3: Results

3.1 Implementation of SAS

Since SAS_v1 and SAS_v2 both use the same data storage and audio-visual playback framework, only data from SAS_v1 are used in Section 3.1 for sanity checks. From a visual and audio standpoint, the implemented SAS application performed in real-time on both ASUS Nexus 7 and Samsung Galaxy Tab E, running on Android API level 19+. The data for each task, both auditory and questionnaire, were also pushed successfully to Parse.com within 10 seconds after the completion of that task. SAS performed in real-time on a Samsung Galaxy Note 7, running on Android API level 21. However, while the program was computationally smooth, the size of the buttons and letters were significantly scaled down on the mobile phone device, resulting in a potentially more difficult task.

![Graphs showing Android amplitude presentation levels for different tasks](image)

**Figure 3.1** Android amplitude presentation levels of attend and ignore track for participant 03. Note that although the amplitudes of the individual tracks were varied, the sum, or total Android amplitude, remained constant throughout the task.
Raw data from participant 03’s playthrough was used to perform sanity checks. The Amplitude of Attend Track, Amplitude of Ignore Track and the sum of the two arrays were plotted against individual responses from the subject for all 5 tasks. As expected, the Amplitude of Attend Track decreased as Amplitude of Ignore Track increased. Furthermore, these adjustments in Android amplitudes of individual tracks did not affect the sum of the Android amplitudes.

**Figure 3.2** SNR calculated from recorded Android amplitudes compared to expected outcome. SNR for each response from the subject was calculated from equation 2.3 and the recorded data. Expected SNR change from the one-up-two-down (OUTD) strategy can be seen in red. Notice the difference between the recorded and expected values. This is due to an error in data saving.

The recorded Android amplitudes were used to calculated SNR. The change in SNR was compared against expected SNR changes, calculated from Correctness array. A significant difference was noted in the recorded SNR changes and expected SNR changes. In the case of the subject responding correctly within the time window (correctness = 0), data was incorrectly
saved after the SNR has been updated for the next response. However, data was saved correctly before updating SNR for the next response in both cases when the subject responds too early (correctness = -1) and too late (correctness = 1). This resulted in the recorded amplitude levels to be temporally incorrect. Specifically, in the case when the subject was correct, the recorded amplitudes were at the presentation level of the next response. This discrepancy was only found after human subject data acquisition. Luckily, this was only a data saving error and the actual change in SNR correctly followed the one-up-two-down rule. This was validated through in task real-time printout of SNR. Since the step size for SNR change was set to be 1.25 dB throughout the task, the actual SNR presented to the subject in SAS can be calculated from the Correctness array and the initial SNR conditions of the level. As expected, this newly constructed SNR array fits the expected outcome from the one-up-two-down strategy.

![Figure 3.3](image_url)

**Figure 3.3** Reconstructed SNR compared to expected outcome. SNR for each response was reconstructed from the Correctness array. Expected SNR change from the one-up-two-down (OUTD) strategy can be seen in red. Notice the similarity between the reconstructed and expected values. This corrected for the error in data saving.
As a further sanity check, the percentage of correct responses at task difficulty convergence can be checked against the theoretical value of 0.707 for one-up-two-down psychoacoustic tasks (Levitt 1971). Percentage correct was calculated by using the last 20 responses from the participant, regardless of the target word of the response. The last 20 responses correspond to a point at which the SNR behavior of the subject can be assumed to have converged. This was calculated for each participant and an average correct percentage was calculated for each task.

<table>
<thead>
<tr>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
<th>Task 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>72.9</td>
<td>63.2</td>
<td>63.2</td>
<td>72.1</td>
<td>73.4</td>
</tr>
</tbody>
</table>

Figure 3.4 Average percentage of last 20 responses being correct for different tasks of SAS. The theoretical value at convergence is 70.7%.

The average percentage correct at SAS_v1 convergence is acceptable when compared to the theoretical value. This indicates that the one-up-two-down adaptive strategy is functioning correctly.

### 3.2 SNR Data

SNR data from all subjects was plotted against Word Index of the response word. Due to the nature of SAS, if a subject respond too early to a target word, the SNR will increase but the target word will not change until the subject responds correctly within the timing window or the timing window has passed. Therefore, a subject may have multiple responses to the same target word and the number of subject responses during any particular task varied dramatically between subjects. Thus, the Word Index was chosen as the independent variable to plot SNR data against as the number or chronological order of the target words for a particular task did not change between subjects.
Figure 3.5 SNR at each target word for outliers, data from SAS_v1. Data originates for subject 15 and 27. Note that SNR is strictly increasing in both cases.

During visual inspection of the original plots, the data from subject 15 and 27 appeared to be outliers, with SNR strictly increasing throughout all tasks. Upon closer inspection of the data, both subject 15 and 27 were too late in response for all target words for all levels (Correctness = 1 for all levels). This suggests that the subjects were not responding correctly or may have misunderstood the task. Therefore, the SNR data from subject 15 and 27 were labelled as outliers and discarded for all further analysis.
Figure 3.6 SNR at each target word for SAS_v1 subject data. All subject data except for subject 15 and 27 are shown here.

Similarly, plots for all subjects from SAS_v2 can also be plotted.

Figure 3.7 SNR at each target word for subject data from SAS_v2.
3.3 Questionnaire Data

SAS questionnaire data was processed as described in section 2.3.2. The following are the confusion matrices for the 5 task of SAS. Questionnaire was not included in the protocol for SAS_v2 subjects.

<table>
<thead>
<tr>
<th>Task Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positive</td>
</tr>
<tr>
<td>False Positive</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
<th>Task 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>293</td>
<td>111</td>
<td>256</td>
<td>144</td>
<td>329</td>
</tr>
<tr>
<td>105</td>
<td>159</td>
<td>73</td>
<td>195</td>
<td>78</td>
</tr>
</tbody>
</table>

**Figure 3.8** Confusion matrices for questionnaire answers of each task.

<table>
<thead>
<tr>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
<th>Task 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>QTPR</td>
<td>0.736</td>
<td>0.778</td>
<td>0.808</td>
<td>0.855</td>
</tr>
<tr>
<td>QTNR</td>
<td>0.602</td>
<td>0.728</td>
<td>0.709</td>
<td>0.798</td>
</tr>
<tr>
<td>QFPR</td>
<td>0.398</td>
<td>0.272</td>
<td>0.291</td>
<td>0.202</td>
</tr>
<tr>
<td>QFNR</td>
<td>0.275</td>
<td>0.360</td>
<td>0.178</td>
<td>0.182</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.677</td>
<td>0.675</td>
<td>0.777</td>
<td>0.810</td>
</tr>
</tbody>
</table>

**Figure 3.9** Questionnaire statistics across all participants for all tasks of SAS_v1. Questionnaire true positive rate (QTPR), questionnaire true negative rate (QTNR), questionnaire false positive rate (QFPR), and questionnaire false negative rate (QFNR) as described in section 2.3.2.

As the questionnaire is a modified two alternative forced choice task, the expected accuracy from the subjects is 0.5. The QTPR is above 0.7 for all 5 tasks, suggesting that participants were able to retain the context of the attend conversation after finishing tasks. An interesting observation is that the QTNR is also rather high, suggesting that the participants were also able retain some of the context of the ignore conversation. However, the QTNR is lower than the QTPR for all 5 tasks, further suggesting that participants were less able to associate words with the ignore conversation. While this is expected as the focus of the task is on the attend conversation, the
ability for subjects to retain significant amounts of contextual information from a competing continuous speech stream is poorly documented.

3.4 Linear Models of QuickSIN SNR50 and SAS Metrics
Parallels can be drawn between the mechanics behind QuickSIN and SAS. The scoring metrics (SNR50 and task score) are closely linked to the number of correct responses to words found in speech presentation. In particular, QuickSIN tasks present 6 sentences which contain 5 scored words each. SNR50 is then calculated by subtracting the total number of correctly responded words from 25.5. Similarly, the SAS task SNR difficulty adjusts according to the correctness of responses, and thus the minimum SNR at which a participant can perform is closely related to the total number of correct responses to target words. Intuitively, the prediction of QuickSIN SNR50 is most likely somewhat related to the number of correct responses during SAS tasks. However, since subjects are able to give multiple responses for the same target word in SAS when their first response is too early, the total number of correct first responses during each task would be a more informative and comparable metric when compared to the total number of correct responses to target words, as the number of first responses for the same task across all subjects is the number of target words during the task. Linear models were fit to the data with the methodology described in section 2.3.3.
Figure 3.10 Total number of correct first responses for each task as a linear function of QuickSIN SNR50; data from SAS_v1. Each dot represents a participant from the study. Blue and red dots represent participants classified with QuickSIN as normal hearing and mild speech-in-noise SNR loss respectively.

<table>
<thead>
<tr>
<th>Task</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
<th>Task 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-Squared</td>
<td>0.002</td>
<td>0.053</td>
<td>0.070</td>
<td>0.094</td>
<td>0.125</td>
</tr>
<tr>
<td>Estimate of intercept</td>
<td>-1.399</td>
<td>2.937</td>
<td>3.672</td>
<td>5.328</td>
<td>5.104</td>
</tr>
<tr>
<td>Estimate of slope</td>
<td>0.029</td>
<td>-0.146</td>
<td>-0.162</td>
<td>-0.186</td>
<td>-0.168</td>
</tr>
<tr>
<td>F-statistic for model</td>
<td>0.059</td>
<td>1.795</td>
<td>2.393</td>
<td>3.303</td>
<td>4.580</td>
</tr>
<tr>
<td>P-value for model</td>
<td>0.809</td>
<td>0.190</td>
<td>0.132</td>
<td>0.079</td>
<td>0.040</td>
</tr>
</tbody>
</table>

Figure 3.11 Linear model fit statistics for number of first response correct from different tasks; data from SAS_v1. The F-test was performed as an assessment for fit of model. The null hypothesis of the F-test was that the estimate of the intercept and slope are both 0. Note, in this case, only the model for task 5 is significant at \( p = 0.05 \). However, after corrections, none of the tasks produced an individually significant fit at \( p = 0.01 \).

None of the data from SAS yielded a statistically significant correlated linear model after correction. Furthermore, the R-squared values for all 5 models were relatively low, indicating that the linear regression models provided poor fits between the variables. In order to observe more statistically relevant trends, the total number of correct first responses across all tasks was plotted as a function of QuickSIN SNR50.
Figure 3.12 Comparing total number of correct first responses across all tasks for participants as a function of QuickSIN SNR50; data from SAS_v1.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-Squared</td>
<td>0.074</td>
</tr>
<tr>
<td>Estimate of intercept</td>
<td>4.737</td>
</tr>
<tr>
<td>Estimate of slope</td>
<td>-0.037</td>
</tr>
<tr>
<td>F-statistic for model</td>
<td>2.562</td>
</tr>
<tr>
<td>P-value for model</td>
<td>0.119</td>
</tr>
</tbody>
</table>

Figure 3.13 Linear model fit metrics for total number of correct first responses across all tasks as a function of QuickSIN SNR50; data from SAS_v1. Note, in this case, the model is not significant at p = 0.05.

However, the linear model still is not significant at p = 0.05, and the R-squared value remained small. This is most likely due to the low number of participants with mild SNR loss, resulting in this population becoming a relatively dense cluster in terms of QuickSIN SNR50, therefore limiting the domain and effectiveness of the model.
Similarly, the participant data from SAS_v2 and the corresponding QuickSIN SNR50s can be used to create the following linear models. The data from the 3 participants with QuickSIN SNR50 greater than 6 contributed to this improved significance in linear model fit.

![Figure 3.14](image-url) **Figure 3.14** Total number of correct first responses for each task as a linear function of QuickSIN SNR50; data from SAS_v2. Each dot represents a participant from the study. Blue and red dots represent participants classified with QuickSIN as normal hearing and mild speech-in-noise SNR loss respectively.

![Figure 3.15](image-url) **Figure 3.15** Linear model fit statistics for number of first response correct from different tasks, data from SAS_v2. The F-test was performed as an assessment for fit of model. The null hypothesis of the F-test was that the estimate of the intercept and slope are both 0. Note, in this case, the model for all 6 tasks are significant at p = 0.05. After corrections, all tasks, excluding task 5, produced individually significant fits at p = 0.00833.

Data from all individual tasks of SAS_v2 yielded a significant (p < 0.05) linear model, data from all tasks, excluding task 5, yielded significant models after correction. The R-squared values also
saw improvement. Furthermore, the linear model between the total number of correct first responses and QuickSIN SNR50 was also significant at $p = 0.05$ and had an $R$-squared value of 0.31. This $R$-squared value is within the acceptable range for this study as the human subject data can be classified as stochastic behavioral data.

![SASv2 total correct first responses as a function of QuickSIN SNR50](image.png)

**Figure 3.16** Comparing total number of correct first responses across all tasks for participants as a function of QuickSIN SNR50; data from SAS_v2

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-Squared</td>
<td>0.310</td>
</tr>
<tr>
<td>Estimate of intercept</td>
<td>12.673</td>
</tr>
<tr>
<td>Estimate of slope</td>
<td>-0.056</td>
</tr>
<tr>
<td>F-statistic for model</td>
<td>17.074</td>
</tr>
<tr>
<td>P-value for model</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**Figure 3.17** Linear model fit metrics for total number of correct first responses across all tasks as a function of QuickSIN SNR50; data from SAS_v2. Note, in this case, the model is significant at $p = 0.05$.

Similarly, statistically significant trends also arise from linear models built from the final SAS score of each task and a subject’s QuickSIN SNR50.
Figure 3.18 Final SAS Score for each task as a linear function of QuickSIN SNR50, data from SAS. Each dot represents a participant from the study. Blue and red dots represent participants classified with QuickSIN as normal hearing and mild speech-in-noise SNR loss respectively.

<table>
<thead>
<tr>
<th>Task</th>
<th>R-Squared</th>
<th>Estimate of intercept</th>
<th>Estimate of slope</th>
<th>F-statistic for model</th>
<th>P-value for model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>0.040</td>
<td>0.407</td>
<td>0.000</td>
<td>1.334</td>
<td>0.257</td>
</tr>
<tr>
<td>Task 2</td>
<td>0.144</td>
<td>1.339</td>
<td>0.000</td>
<td>5.396</td>
<td>5.396</td>
</tr>
<tr>
<td>Task 3</td>
<td>0.148</td>
<td>0.808</td>
<td>0.000</td>
<td>5.554</td>
<td>5.554</td>
</tr>
<tr>
<td>Task 4</td>
<td>0.082</td>
<td>1.073</td>
<td>0.000</td>
<td>2.846</td>
<td>2.846</td>
</tr>
<tr>
<td>Task 5</td>
<td>0.208</td>
<td>1.250</td>
<td>0.000</td>
<td>8.401</td>
<td>8.401</td>
</tr>
</tbody>
</table>

Figure 3.19 Linear model fit metrics for final SAS score; data from SAS_v1. The F-test was performed as an assessment for fit of model. The null hypothesis of the F-test was that the estimate of the intercept and slope are both 0. Note, in this case, fits for tasks 2, 3 and 5 are significant at $p = 0.05$. However, after corrections, only task 5 produced an individually significant fit at $p = 0.01$. 
Figure 3.20 Comparing total final SAS scores across all tasks for participants as a function of QuickSIN SNR50; data from SAS_v1.

Figure 3.21 Linear model fit metrics for total final SAS scores across all tasks as a function of QuickSIN SNR50; data from SAS_v1. Note, in this case, the model is significant at p = 0.05.
Figure 3.22 Final SAS Score for each task as a linear function of QuickSIN SNR50; data from SAS_v2. Each dot represents a participant from the study. Blue and red dots represent participants classified with QuickSIN as normal hearing and mild speech-in-noise SNR loss respectively.

<table>
<thead>
<tr>
<th>Task</th>
<th>R-Squared</th>
<th>Estimate of intercept</th>
<th>Estimate of slope</th>
<th>F-statistic for model</th>
<th>P-value for model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>0.227</td>
<td>2.805</td>
<td>0.000</td>
<td>11.186</td>
<td>0.002</td>
</tr>
<tr>
<td>Task 2</td>
<td>0.224</td>
<td>6.217</td>
<td>0.000</td>
<td>10.957</td>
<td>0.002</td>
</tr>
<tr>
<td>Task 3</td>
<td>0.423</td>
<td>4.766</td>
<td>0.000</td>
<td>27.888</td>
<td>0.000</td>
</tr>
<tr>
<td>Task 4</td>
<td>0.268</td>
<td>5.984</td>
<td>0.000</td>
<td>13.923</td>
<td>0.011</td>
</tr>
<tr>
<td>Task 5</td>
<td>0.129</td>
<td>3.225</td>
<td>0.000</td>
<td>5.644</td>
<td>0.023</td>
</tr>
<tr>
<td>Task 6</td>
<td>0.362</td>
<td>5.202</td>
<td>0.000</td>
<td>21.548</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Figure 3.23 Linear model fit metrics for final SAS score; data from SAS_v2. The F-test was performed as an assessment for fit of model. The null hypothesis of the F-test was that the estimate of the intercept and slope are both 0. Note, in this case, fits for all tasks are significant at p = 0.05. After corrections, all tasks, excluding task 5, produced individually significant fits at p = 0.00833.
Figure 3.24 Comparing total final SAS scores across all tasks for participants as a function of QuickSIN SNR50; data from SAS_v2.

Figure 3.25 Linear model fit metrics for total final SAS scores across all tasks as a function of QuickSIN SNR50, data from SAS_v2. Note, in this case, the model is significant at p = 0.05.

Since SAS score is dependent on the number of correct responses that the subject gave throughout the task, it is expected to have similar linear model fit behavior when compared to the total number of correct first responses. However, due to the design of the scoring system, the final score for each participant is also related to the SNR at which subjects performed throughout the task. The score is an aggregate metric which encodes for more information than correctness alone. Therefore, the score is expected to be a better explanatory variable, resulting in more tasks
having significant fits and higher R-squared values when compared to that fits number of correct responses.

**Figure 3.26** Minimum SAS SNR for each task as a linear function of QuickSIN SNR50; data from SAS_v1. Each dot represents a participant from the study. Blue and red dots represent participants classified with QuickSIN as normal hearing and mild speech-in-noise SNR loss respectively.

<table>
<thead>
<tr>
<th>Task</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
<th>Task 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-Squared</td>
<td>0.030</td>
<td>0.223</td>
<td>0.145</td>
<td>0.059</td>
<td>0.153</td>
</tr>
<tr>
<td>Estimate of intercept</td>
<td>-0.303</td>
<td>0.487</td>
<td>-0.002</td>
<td>0.136</td>
<td>0.124</td>
</tr>
<tr>
<td>Estimate of slope</td>
<td>0.052</td>
<td>0.162</td>
<td>0.093</td>
<td>0.069</td>
<td>0.095</td>
</tr>
<tr>
<td>F-statistic for model</td>
<td>0.986</td>
<td>9.192</td>
<td>5.442</td>
<td>2.000</td>
<td>5.773</td>
</tr>
<tr>
<td>P-value for model</td>
<td>0.328</td>
<td>0.005</td>
<td>0.026</td>
<td>0.167</td>
<td>0.022</td>
</tr>
</tbody>
</table>

**Figure 3.27** Linear model fit metrics for minimum SAS SNR, data from SAS. The F-test was performed as an assessment for fit of model. The null hypothesis of the F-test was that the estimate of the intercept and slope are both 0. Note, in this case, fits for tasks 2, 3 and 5 are significant at p = 0.05. However, after corrections, only task 2 produced an individually significant fit at p = 0.01.
Figure 3.28 Comparing total minimum SAS SNR across all tasks for participants as a function of QuickSIN SNR50, data from SAS.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-Squared</td>
<td>0.139</td>
</tr>
<tr>
<td>Estimate of intercept</td>
<td>0.274</td>
</tr>
<tr>
<td>Estimate of slope</td>
<td>0.023</td>
</tr>
<tr>
<td>F-statistic for model</td>
<td>5.149</td>
</tr>
<tr>
<td>P-value for model</td>
<td>0.030</td>
</tr>
</tbody>
</table>

Figure 3.29 Linear model fit metrics for total minimum SAS SNR across all tasks as a function of QuickSIN SNR50, data from SAS. Note, in this case, the model is significant at $p = 0.05$. 
Figure 3.30 Minimum SAS SNR for each task as a linear function of QuickSIN SNR50, data from SAS_v2. Each dot represents a participant from the study. Blue and red dots represent participants classified with QuickSIN as normal hearing and mild speech-in-noise SNR loss respectively.

<table>
<thead>
<tr>
<th>Task</th>
<th>R-Squared</th>
<th>Estimate of intercept</th>
<th>Estimate of slope</th>
<th>F-statistic for model</th>
<th>P-value for model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>0.229</td>
<td>2.197</td>
<td>0.076</td>
<td>11.281</td>
<td>0.002</td>
</tr>
<tr>
<td>Task 2</td>
<td>0.283</td>
<td>4.240</td>
<td>0.106</td>
<td>15.030</td>
<td>0.000</td>
</tr>
<tr>
<td>Task 3</td>
<td>0.396</td>
<td>3.443</td>
<td>0.133</td>
<td>24.956</td>
<td>0.000</td>
</tr>
<tr>
<td>Task 4</td>
<td>0.238</td>
<td>3.925</td>
<td>0.092</td>
<td>11.853</td>
<td>0.001</td>
</tr>
<tr>
<td>Task 5</td>
<td>0.198</td>
<td>2.466</td>
<td>0.119</td>
<td>9.390</td>
<td>0.004</td>
</tr>
<tr>
<td>Task 6</td>
<td>0.341</td>
<td>3.374</td>
<td>0.196</td>
<td>19.647</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Figure 3.31 Linear model fit metrics for minimum SAS SNR, data from SAS_v2. The F-test was performed as an assessment for fit of model. The null hypothesis of the F-test was that the estimate of the intercept and slope are both 0. Note, in this case, fits for all tasks are significant at p = 0.05. After corrections, all tasks produced individually significant fits at p = 0.00833.

<table>
<thead>
<tr>
<th>Task</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
<th>Task 5</th>
<th>Task 6</th>
</tr>
</thead>
</table>

Figure 3.32: Average SAS minimum SNR differences between normal and impaired. Note that in this case, impaired is different from the mildly impaired definition used for plots, and is defined to be 6+ dB QuickSIN SNR50.
Figure 3.33 Comparing total minimum SAS SNR across all tasks for participants as a function of QuickSIN SNR50, data from SAS_v2.

R-Squared | 0.357
---|---
Estimate of intercept | 3.792
Estimate of slope | 0.024
F-statistic for model | 21.130
P-value for model | 0.000

Figure 3.34 Linear model fit metrics for total minimum SAS SNR across all tasks as a function of QuickSIN SNR50, data from SAS_v2. Note, in this case, the model is significant at $p = 0.05$.

Similar significant linear models also arise from minimum SAS SNR of each task, suggesting that SAS is indeed able to assess a participant’s speech-in-noise hearing ability similar to that of QuickSIN. It is interesting to note that, as explanatory variables, both SAS score and minimum SNR produced more statistically significant fits when compared to number correct first responses. It is also interesting to note that the data from SAS_v2 seem to produce more statistically significant models overall across all 3 SAS metrics used. This can be mainly attributed to the 3 participants for SAS_v2 with higher QuickSIN SNR50 (>6 dB) effectively
increasing the range of the models. This suggests that a larger participant group with a larger variation in QuickSIN SNR50 will be helpful to establish more definitive trends.

3.5 Internal Consistency

Below are the Cronbach’s Alphas calculated with final SAS score and minimum task SNR as the two different metrics.

<table>
<thead>
<tr>
<th></th>
<th>SAS_v1</th>
<th></th>
<th>SAS_v2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SAS Score</td>
<td>Minimum SNR</td>
<td>SAS Score</td>
<td>Minimum SNR</td>
</tr>
<tr>
<td>Sum of Across Tasks</td>
<td>8.8E+08</td>
<td>589.040</td>
<td>2.19E+09</td>
<td>2247.69</td>
</tr>
<tr>
<td>Variance Associated with Total Scores</td>
<td>2.26E+08</td>
<td>149.930</td>
<td>5.67E+08</td>
<td>516.91</td>
</tr>
<tr>
<td>Cronbach's Alpha</td>
<td>0.929</td>
<td>0.932</td>
<td>0.889</td>
<td>0.924</td>
</tr>
</tbody>
</table>

Note that the Cronbach’s alphas are both larger than 0.85 and lower than 0.95, which is above the acceptable alpha threshold of 0.7. This indicates that the individual tasks are relatively reliable measures for both SAS score and minimum SAS SNR. However, the alphas are somewhat above the recommended maximum threshold of 0.9, suggesting that some aspects of the tasks may be redundant and the tasks can be lowered in length (Tavakol and Dennick 2011).

The inter-task correlation can also be found by performing linear regression on the same metric from different tasks. Below are the inter-task correlation plots for SAS score and minimum SNR performance for the SAS participant group.
Figure 3.36 Inter-task correlation for SAS Score; data from SAS_v1.

Figure 3.37 Inter-task correlation for minimum SAS SNR performance; data from SAS_v1.
Overall, the inter-task Pearson correlation coefficients for the two metrics are high for all task to task comparisons. While the underlying metrics are stochastic in nature, these high correlation coefficients suggest that the individual tasks are measuring similar metrics. Similar trends can also be shown with the SAS_v2 dataset.

Figure 3.38 Inter-task correlation for SAS Score; data from SAS_v2.
The inter-task correlation is lower for some combination of tasks in the SAS_v2 dataset. For example, the Pearson correlation coefficient for SAS score between tasks 1 and 2 is only 0.48, suggesting a rather poor fit for the linear model. This increased deviation can be attributed to the dichotic nature of the tasks. In task 1 of SAS_v2, the attend track is presented in the left ear, while the attend track is presented to the right ear in task 2. These changes in both attending ear and conversation resulted in the smaller Pearson correlation coefficient. However, the overall inter-task correlation remains high for a behavioral audiometric task where the underlying function is stochastic.
Chapter 4: Discussion

4.1 Viability of Hearing Tests on Android Devices
Numerous audiometric tests have been developed for and tested on mobile devices in the past 10 years. However, much of the effort has been devoted to mirroring traditional pure tone clinical tests onto mobile platforms (Abu-Ghanem and Handzel 2015). Pure tone audiometry is a method of obtaining hearing threshold levels of an individual through using presenting sinusoidal waveforms of different frequencies and intensities. Many of these mobile pure tone tasks can measure pure tone thresholds in a quiet room or sound booth, similar to clinically administered pure tone tests (Peer and Fagan 2015, Saliba, Al-Reefi et al. 2017). While SAS is a speech-in-noise audiometric task which differs in nature from traditional pure tone audiometry, it faces similar obstacles in testing setup.

Android-capable mobile devices are able to playback multiple sound sources through its auditory auxiliary port for extended periods of time, fulfilling the basic hardware requirements for an effective hearing test. This port is capable of delivering different sounds through to the left and right channels at the same time, allowing for additional flexibility in audiometric testing. Furthermore, most Android-capable devices are also internet capable, allowing for centralized collection and automated analysis of user generated data. This is an improvement from traditional clinical tests, contributing to a potentially large database of audiology data for research and freeing up valuable clinical time from audiologists.

However, mobile audiometric procedures still face some hardware obstacles. Firstly, different models of mobile devices contain different audio playback hardware on logic boards, resulting in differences in frequency response and audio power. A difference in headphones used during the
task can create further variations in testing conditions, making it harder to have valid comparisons between data from different users. In fact, headphone acoustics have been noted to affect the quality of data collected and be extremely important in mobile hearing test applications (Pickens, Robertson et al. 2017). While these obstacles were bypassed in this study by standardizing the testing hardware, such differences in user hardware will be prevalent in the real world. However, these obstacles can be potentially mitigated through smart test design and testing procedures. For example, a speech-in-noise audometric test attempts to measure a signal-to-noise ratio as a threshold rather than the intensity threshold measured by pure tone tests. The ratio threshold will mitigate some of the needs for precision hardware as both signal and noise are delivered through the same system, so some of the systematic errors maybe cancelled through division. Such systematic errors may be a prevalent problem in mobile pure tone tests. A potential calibration method would be to ask the user for device models before administering the test and comparing the generated data against data from users with similar hardware setups.

In terms of software, Android devices are capable of delivering sounds at a variety of amplitudes that can be configured in real-time. The default MediaPlayer class for Android audio playback takes in two numbers between 0 and 1 as the playback amplitudes in the left channel and right channels respectively. These numbers are linearly related to the voltage amplitude of the audio output, allowing for easy audio intensity manipulations. In terms of input audio data, the Android platform can supports a variety of formats, such as WAVE, MP3, ACC and FLAC, at different sampling rates. This allows for playback of most digital audio recordings. Overall, mobile devices on the Android platform forgo some of the precision offered by clinical audimetric equipment but retain all of the basic requirements. However, mobile devices also offer data
storage and playback capabilities beyond that offered by traditional equipment. This makes mobile devices especially suitable for speech-in-noise audiometric testing purposes.

4.2 Effectiveness of Semantic Auditory Search
In retrospect, the original design for SAS shown in figure 2.1 had multiple continuous speech noise sources and additional spatial components which made it difficult to compare to traditional speech-in-noise audiometric tasks. Being a novel proposed audiometric task, it was hard to draw direct parallels between the original design and current clinical speech-in-noise audiometric tests. However, the reduced forms of SAS_v1 and SAS_v2 presented to subjects in this study is more akin to clinical speech-in-noise tasks in task delivery, and is the user data more suited to be compared against that from QuickSIN.

In this study, a variety of SAS metrics were benchmarked against QuickSIN’s SNR50 in order to validate SAS as a speech-in-noise audiometric task. From the simple linear regression models found in section 3.4 it can concluded that all three of the SAS metrics surveyed: number of correct first responses, SAS score and minimum SNR of performance, were somewhat related to the subject’s SNR50 measured through QuickSIN. In terms of performance, SAS score and minimum SNR of performance were better explanatory variables for QuickSIN SNR50 when compared to number of correct first responses. As expected, the total of metrics across all tasks for SAS_v1 or SAS_v2 consistently produce models with higher goodness of fit r-squares than the same metrics for a single task.

Total metric linear fits from figure 3.24 and figure 3.34 demonstrate that SAS score and minimum SNR of performance metrics from SAS_v2 are statistically significant when linearly correlated with QuickSIN SNR50 values at a significance level of p = 0.05. The r-squared values
of 0.368 and 0.357 also indicate that these metrics account for a substantial portion of the variance found in the QuickSIN SNR50 data. In biomedical sciences correlation coefficients larger than 0.3 suggest a medium effect of the explanatory variables on the response variable (Cohen 1988). This effect is significant even with relatively small Pearson coefficient values as much of human behavior, a cornerstone in audiometric testing, is often stochastic. This underlying stochasticity introduces additional variance into the data, resulting in a relatively poor fit for the linear regression models. While similarly high values of r-squared do not hold for the same metrics in figure 3.20 and figure 3.28, it can be mainly attributed to the limited distribution of QuickSIN SNR50 found in the SAS_v1 participant group, from -3dB to 3dB, compared to that of the SAS_v2 participant group, from -2dB to 8dB. Overall, the linear model statistics from section 3.4 suggests that user performance in SAS is related to performances in QuickSIN, which indicates that SAS is a somewhat effective speech-in-noise audiometric test. Another method to validate SAS as a speech-in-noise task is to observe the effects of speech intelligibility in competing conversation. On average, for a single competing talker, the expected increase in SNR50 is estimated to be approximately 10-15 dB (Carhart and Tillman 1970, Festen and Plomp 1990). The data from figure 3.30 is consistent with that found in literature, with the mean of the mild hearing impaired population approximately 9-17 dB above that for the normal hearing population throughout all 6 of the tasks, as seen in figure 3.32. While SAS may lack some of the clinical validation that other speech-in-noise tests such as QuickSIN have, it also offers many improvements on top of traditional audiometric tests. The Android mobile platform allows SAS to be far more accessible, this allows SAS to become a discrete, automated screening task for hearing impairment. SAS also offers a more realistic auditory testing environment when compared to traditional tasks. An important difference between SAS and QuickSIN is that SAS
presents contextually rich conversations instead of simple phonetically balanced sentences found in QuickSIN. In a sense, SAS takes into account the subject’s listening ability as well as hearing ability.

This distinct contextual difference between SAS and QuickSIN signal streams also gives rise to the difference between QuickSIN SNR50 and minimum SAS SNR performance. In figure 3.30, the QuickSIN SNR50 ranges from -10dB to 5dB, while the minimum SAS SNR performance ranges from approximately -30dB to 10dB. These observations are consistent with the reported improved word recognition in contextually rich sentences when compared to sentences embedded in less context (Bilger, Nuetzel et al. 1984).

SAS can also be considered as an automated audiometric testing platform. The software package was designed so that new tasks can be made by providing an attend track, an ignore track and a target word list. This potentially allows users to select their favorite podcasts, audiobooks or other speech recordings as the attend stream. This allows for potential long term audiometric assessments, with automated data collection, which could provide new insights. As seen in the case for SAS_v2, the framework can be modified to deliver both dichotic as well as diotic tasks. Furthermore, the scoring system could serve alongside reimbursement incentives to create a multitude of psychometric tests to gauge motivation.

An interesting, but expected observation is that the number of user who presumably misunderstood the task and had no task input, as shown in figure 3.5, decreased from 2 to 0 when comparing SAS to SAS_v2. While both participant groups were given similar instructions and performed under similar environmental conditions, the user interface of the tasks changed as seen in both figure 2.2 and 2.7. This simplification of the task’s graphic user interface likely
improved general ease of use and should be considered as a design principle for mobile audiometric tasks in the future.

4.3 Future Directions
While this study is able to establish SAS as a potential speech-in-noise audiometric screening task, many of SAS’s additional features are poorly explored. Perhaps the most obvious route for future research would be to reproduce the results of this study in a larger study population. While 79 subjects participated in both SAS_v1 and SAS_v2, the QuickSIN SNR50 distribution for the participants was rather compact. Luckily, SAS is a testing platform that can be easily distributed on the Android platform, allowing for remote participation and data collection. However, in-app delivery of current protocols and instructions need to occur before the task can be truly automated and widely distributed.

In terms of data analysis, this study focused heavily on regression analysis as classification performs rather poorly when the data set is imbalanced. However, classification remains an important feature for speech-in-noise tasks as discretely classified groups allows for easier diagnostic decisions. For SAS_v1, the subject numbers ratio of mild hearing loss to normal hearing participants was 2:35, while for SAS_v2, the ratio was 6:36. Therefore, it would be prudent to carry out classification experiments after accumulating more data.

Furthermore, it would be prudent to perform deeper analysis of the SAS platform. For example, the relationship between scoring speech-in-noise tasks per word and per phoneme can be easily explored through SAS, as the target word lists, and therefore targeted phonemes, for a particular signal track is standardized across all participants.
Ultimately, this thesis has proven that the Semantic Auditory Search framework is a scalable, automated speech-in-noise audiometric task that can potentially replace current clinical tests. Furthermore, the Android mobile platform is capable of facilitating largescale, distributable audiometric testing.
References


Appendix A: 
Transcripts of Audio Tracks used in SAS

The following are the transcripts of the audio tracks used in for the tasks found in SAS. The numbering for the tracks correspond to those found in figure 2.6. All tracks were sourced from National Public Radio.

Track 1

MELISSA BLOCK, HOST:

This is ALL THINGS CONSIDERED from NPR News. I'm Melissa Block.

It's a sign of deeply partisan times when a Democratic president and a Republican House speaker make headlines just by sitting down and talking to each other. That's what happened today in a rare hour-long meeting that aides call constructive. How constructive is not exactly clear. And while the president and House speaker agreed to work together in areas where there's common ground, that appears to be very small territory.

NPR White House correspondent Scott Horsley joins me now. And, Scott, this was the first advertised one-on-one meeting between President Obama and House Speaker John Boehner in 14 months. What took them so long?

SCOTT HORSLEY, BYLINE: Well, you can ask what too so long, or you could just as easily ask why they bothered to meet today. Most of their earlier meetings have come when there was some sort of fiscal deadline looming. But today, there's no such forcing mechanism. We're not looking at a government shutdown or a breach of the debt limit. And what's more, for the past few weeks, the president has been largely ignoring lawmakers, trying to do what he can through executive action. Nevertheless, White House spokesman Jay Carney said this afternoon the president is not giving up on Congress.

JAY CARNEY: We're going to continue to engage with Congress, with Republicans in an effort to see where we can find common ground and move the ball forward for the American people. Where Congress refuses to act, the president is going to use every authority available to him to advance an agenda that expands opportunity and rewards hard work.

HORSLEY: And while this was the first time in over a year that a meeting like this showed up on the president's schedule, the White House hints there might have been conversations and meetings that weren't made public. Certainly, Obama and Boehner have been together as part of bigger negotiations, for example, during last fall's government shutdown.
BLOCK: So what did President Obama and Speaker Boehner talk about? And did they end up finding any of that common ground that Jay Carney was talking about?

HORSLEY: Well, they covered a lot of ground, I'm not sure how much of it was common. According to the speaker's office, they talked about trade, immigration, drought relief, and the president's health care law, among other things. Now, Republicans are actually more supportive of the president's trade agenda than a lot of Democrats are. But otherwise, there's not a lot of room where you can see agreement here. Immigration looks to be stalled in the Republican House.

On the drought, which is certainly affecting both Republicans and Democrats out West, the two parties have very different ideas about how to deal with it. Obama has threatened to veto the House Republican plan. And, of course, health care remains the big divide. Barely an hour after leaving the Oval Office, Speaker Boehner was on the House floor denouncing Obamacare, saying it would drive up insurance premiums for small businesses.

REPRESENTATIVE JOHN BOEHNER: Another sucker punch to our economy, another broken promise to hardworking Americans. That's why the House continues to focus on stopping government abuse and promoting better solutions for middle-class families and small businesses.

HORSLEY: So just in case you thought they were going to be making nice, think again. As Jay Carney said, it was just a meeting.

BLOCK: Just a meeting. Well, Scott, next week, the president is going to release his budget proposal. And I gather the White House has already telegraphed that it is not going to include the olive branch to Republicans that the president offered last year on entitlement programs such as Social Security.

HORSLEY: That's right. Last year, the White House budget included a proposal to change the way the government makes cost of living adjustments in Social Security and lots of other programs.

BLOCK: Was this chained CPI we're talking about again?

HORSLEY: That's right.

BLOCK: OK.

HORSLEY: And, you know, it would have both reduced government spending and also boosted revenue. A lot of Democrats hated the idea. But it was designed as a way to reach out to Republicans, bring them to the negotiating table in hopes of striking a grand budget bargain. And this was coupled with a White House charm offensive, lots of dinners and coffees. But the president wanted to see additional revenue as well, and that was a deal breaker for the GOP. So it
wound up going nowhere. And this year's budget includes no such olive branch. It's going to be instead a much more partisan Democratic budget.

BLOCK: OK. NPR White House correspondent Scott Horsley. Scott, thanks so much.

HORSLEY: My pleasure.

Track 2

RENEE MONTAGNE, HOST:

It's MORNING EDITION, from NPR News. Good morning. I'm Renee Montagne.

DAVID GREENE, HOST:

And I'm David Greene.

On the Ukrainian peninsula of Crimea, armed men, widely believed to be Russian soldiers, are firmly in control. This military standoff has been peaceful so far, a relief to the United States and its allies who are seeking a diplomatic resolution with Russia.

MONTAGNE: Yesterday, the German and British leaders got on the phone with Russia's President Vladimir Putin. Ukraine's prime minister will meet President Obama in Washington later this week. But so far, there are few signs that these efforts are working.

GREENE: NPR's Emily Harris has been reporting from Crimea, and she has now moved to the Ukrainian capital, Kiev, and joins us on the line. Emily, good morning.

EMILY HARRIS, BYLINE: Good morning, David.

GREENE: And so let's start with some of the phone calls that we've been hearing about. Germany's Chancellor Angela Merkel, British Prime Minister David Cameron both have been on the phone with Vladimir Putin. Do you get a sense of the gist of these conversations?

HARRIS: From the statements that the sides are putting out, they all say they want to deescalate the tensions in Crimea. But the Kremlin says that Putin, in his conservations, pointed out differences that they have on how to see this conflict, in particular, over who has the right, basically, to be in charge. Russia supports the current leadership in Crimea as legitimate. Europe and the U.S. completely disagree with that.

The Kremlin also says that in the phone conversations, Putin was criticizing the current authorities in Kiev for not reining in what Russia calls ultranationalist Ukrainian groups across the country. Essentially, Merkel and Cameron have been trying to get President Putin to at least agree to set up a group to talk about the crisis, and Putin said he would discuss this idea today with his advisors.
GREENE: So, a lot of disagreement remaining. And any signs at all pointing to some sort of resolution at this point?

HARRIS: There has been some contact reported over the past week between Kiev and Moscow, direct contact, which is how the U.S. and Europe say this ultimately has to be resolved, but no progress from there. Putin keeps talking about what's known as the February 21st Agreement, which was a deal between then-President Yanukovych of Ukraine and opposition leaders. It was signed the day after of the worst violence in Kiev, in which scores of people were shot and killed. That agreement called for presidential elections by December and establishing a government of, quote-unquote, "national trust," a lot of political water under the bridge since then. But Moscow still seems to want to somehow make sure that its interests or the interests of the ethnic Russian population in Ukraine will be well represented in Kiev.

GREENE: So, Emily, I mean, even if this new government in Kiev is not sticking to that February 21st Agreement that Putin is talking about, there are plans for presidential elections underway already in Ukraine. Couldn't that be some sort of fresh start?

HARRIS: The U.S. says that's the next step, but Russia says that the current government in Kiev was put in place through a coup, so whatever it does is not legitimate. It will be important to see who runs and what the results are, and that may affect the Kremlin's reactions. This weekend, the leader of a Ukrainian nationalist coalition, the Right Sector, which played a significant role in pushing the old government out, he announced plans to run. Former Prime Minister Yulia Tymoshenko could potentially run. She has had good relations with Putin in the past.

There's a strong desire among a lot of Ukrainians to have a completely fresh start, as you put it, a feeling that many politicians have, in some way, tainted by corruption when you've been in office. But before we get to May, which is a couple months away, the election that everyone's got their eye on now is next Sunday in Crimea, where the current authorities there are trying to put together a vote asking residents if they would like to join Russia. The West says this vote is illegal and won't recognize the results, especially if it's pro-Russia. But Moscow could potentially use those results to harden its position.

GREENE: And Emily, you've just been reporting in Crimea, and have just left. What were some of your impressions about the situation?

HARRIS: I watched the pro-Russia Crimean prime minister swear in a few dozen people into what he called the beginning of the Crimean Armed Forces. I went up to the edge of Crimea, where it rejoins the Ukrainian mainland, and saw Russian soldiers dug in with sandbags, checking people driving through, and well-armed men in uniforms with no insignia. They've clearly set up a camp there.

There haven't been any serious confrontations between the Russian troops and Ukrainian troops, but there have been some conflicts between demonstrators and some of Crimea's Muslim...
population. The Crimean Tatars have reported getting Xs marked on the doors of their homes. This has significant, frightening historical resonance from the time when the Tatars were deported from Crimea under Soviet rule. So one big question is: Will something flare up, either between populations or militaries, in Crimea before diplomacy can work?

NPR's Emily Harris, joining us from the Ukrainian capital, Kiev. Emily, thanks very much.

Thanks, David.

Track 3

LINDA WERTHEIMER, HOST:

Some of the nation's largest pharmaceutical companies have slashed the payments they give to doctors and other health professionals for promoting their products. That's the finding of a new investigation by the nonprofit newsroom ProPublica. The decline in spending comes as more companies have voluntarily posted what they pay doctors who promote their drugs. And it's happening as a deadline approaches for mandatory disclosure.

Charles Ornstein of ProPublica led the investigation and he joins us now. He's in our studio in New York. Good morning.

CHARLES ORNSTEIN: Thanks for having me.

WERTHEIMER: Now, first of all, could you just kind of step back a bit and tell us the background here? How do these payments work? What do the companies want these doctors to do for them? What do they think they're getting?

ORNSTEIN: Well, for many decades, we've known almost nothing about this, but pharmaceutical companies have long worked with physicians and other health professionals to educate their peers about drugs and to help promote sales of their products. Some doctors can earn hundreds of thousands of dollars a year for speaking engagements.

Pharmaceutical companies hope this increases the, you know, sales of their products and gets them into new markets and new physician hands.

WERTHEIMER: Should the fact that docs are flogging drugs for cash make us suspicious or should we assume that they're just simply spreading the word about some good drug they routinely prescribe?

ORNSTEIN: Well, in recent years, this has been incredibly controversial. A number of pharmaceutical companies have paid huge sums, billions of dollars in some cases, to settle whistleblower lawsuits alleging that they improperly marketed their products and paid kickbacks to physicians in exchange for them prescribing their products. So I think that this has been looked upon less favorably in recent years in a number of major universities, and academic
medical centers have gone so far as to prohibit their faculty members from giving these paid talks for drug companies.

WERTHEIMER: Now, your new analysis shows that the payments have started to slack off. In one case, Eli Lilly, which is a very big manufacturer of psychiatric drugs, Lilly cut spending in half from 2011 to 2012. Could you tell us more about the kind of thing you found?

ORNSTEIN: Well, Lilly was just one example. The drug company Pfizer, which is another huge company, dropped by more than 60 percent and the world's biggest drug company in terms of U.S. sales, Novartis, they cut their payments by 40 percent. So I think that there's a couple things that are going on here. One, is these companies have some big blockbuster drugs that are losing their patents and so they're facing generic competition.

And when that happens, the companies tend to pare back spending on promoting those drugs because they're not going to be prescribed anymore. So that's going on. But on the other hand, we're also seeing this wave of transparency and as patients are able to look up and see how much money their physicians earn by working with the pharmaceutical industry, we're seeing, certainly, that the payments are going down.

WERTHEIMER: I understand the Affordable Care Act requires companies to disclose all such payments and that will start in September. Do you imagine that just shining a little light on the subject will cause these kinds of payments to drop still more?

ORNSTEIN: Well, most people think about Obamacare in terms of the health exchanges or Medicaid expansion and that's where the real focus has been, but I think this provision within the act called the Sunshine Act, has the potential to really change things up quite a bit because when all patients are able to look up their physicians and see their interactions with the pharmaceutical industry, I expect there to be a lot more give and take and questions that patients ask of their physicians.

And that could have the effect of a physician saying, you know what, it's not worth it. I don't want to have a relationship if I'm going to on a website and my patient's going to come in and instead of asking me about their health condition, they're going to be asking me about my financial relationship.

WERTHEIMER: It's been our experience in commerce of all kinds that if you find that something stops working for you, then you do something else. So if paying docs to pump up the reputation of drugs, if that's dwindling, what's next for these big drug companies? What are they going to do?

ORNSTEIN: Well, that's an excellent question. Drug companies have a lot of ways of working with physicians and other health professionals. So there are research funds. They pay them to serve on advisory boards. But also the Affordable Care Act, the Sunshine Act, does not require
drug companies to disclose how much they provide to, say, physician assistants or nurse practitioners, and there's a number of folks who think that some of the funding will shift to those practitioners because it doesn't have to be disclosed.

WERTHEIMER: Charles Ornstein is a senior reporter for ProPublica. He joined us from our studios in New York City. Thank you very much.

ORNSTEIN: Thanks for having me.

Track 4

MELISSA BLOCK: This is ALL THINGS CONSIDERED from NPR News. I'm Melissa Block.

AUDIE CORNISH, HOST:

And I'm Audie Cornish.

Yesterday, he was in Ukraine. Today, it's Paris. Secretary of State John Kerry is logging a lot of miles these days, trying to find a diplomatic solution to the unfolding crisis in Crimea. In Paris, he met with Russia's foreign minister. Kerry said the discussions were substantive. Diplomatic sources tell NPR the results were inconclusive. But at least the talking had begun.

Our own Michele Kelemen is traveling with the secretary and joins us from Paris. And, Michele, inconclusive talks between Kerry and the Russian foreign minister, but most of the attention in Paris seems to be on a meeting that didn't happen. Tell us about it.

MICHELE KELEMEN, BYLINE: That's right, Audie. You know, Kerry brought Ukraine's foreign minister, Andrei Deshtitsa, from Kiev to Paris, and all day seemed to be trying to get him together with Lavrov. The first attempt was when Kerry called a meeting of the countries that signed this non-intervention pact with Ukraine back in 1994 - that's when Ukraine gave up that Soviet-era nuclear stockpile. And Kerry opened the meeting at the U.S. ambassador's residence saying one country was regrettably not fair. That, of course, was Russia.

Kerry later went into his meeting with Sergei Lavrov at Russia's foreign - at the Russian ambassador's residence here and then both of them went to this hastily arranged meeting at the French foreign ministry on Ukraine. We watched Lavrov walk in. And then, soon after, Ukraine's foreign minister, who we got to know on the plane last night, was led upstairs while Lavrov and the others were downstairs.

And we kept waiting for some sort of diplomatic choreography for the two to meet. After all, they were in the same building. But it didn't happen. Lavrov left. And when we ask him if he met his Ukrainian counterpart, he said, who's that?

CORNISH: So what...
KELEMEN: You know, the Russians don't recognize this new government. And then when Deshtitsa left, he shrugged and said we'd have to ask Lavrov why they didn't meet.

CORNISH: Meanwhile, what did Secretary Kerry have to say? How did he explain this failure to bring the Russians and Ukrainians together?

KELEMEN: Yeah. He claimed that he had no expectation of a meeting, that he brought the Ukrainian foreign minister here because he wanted to be able to consult him and not just talk to Western European foreign ministers and Lavrov about Ukraine. He said it would have been inappropriate to make decisions with others about Ukraine without consulting Ukrainians. But, you know, having been on the plane with the foreign minister, we know certainly that he had that expectation.

CORNISH: The Russians and the U.S. have very different visions of what's been happening in Ukraine. And we know they're talking, but are they essentially talking past each other?

KELEMEN: It's interesting, Audie. You know, both Lavrov and Kerry came out of their meetings tonight using exactly the same phrase, saying they both wanted to normalize and stabilize the situation and overcome the crisis in Ukraine. But, really, they do seem to be talking past each other on what's actually happening.

The U.S. has been calling on Russia to pull its troops back to barracks in Crimea. Lavrov complains that - he claims that Russian troops are their bases and that they've just taken extra security measures. And he says Russia can't give orders to those who he calls self-defense forces in Crimea.

CORNISH: Michele, do you get the sense that the U.S. is worried that what's happened in Crimea could act as a model for Russian behavior in other contentious areas?

KELEMEN: Well, you know, there is a concern that this rift between the U.S. and Russia now could have an impact on other issues. But on the other hand, Kerry has been sounding increasingly annoyed with Russia when it comes to Syria, for instance, saying that Russia has been increasing its support for the Assad regime. And, of course, the peace talks are going nowhere. He met in Paris today with the international envoy on Syria, Lakhdar Brahimi, but neither man would say anything when I asked them if this dispute over Ukraine is affecting that diplomacy.

CORNISH: That's NPR's Michele Kelemen in Paris, traveling with Secretary of State John Kerry. Michele, thank you.

KELEMEN: Thank you, Audie.

Track 5
MELISSA BLOCK, HOST:
We're going to cut through some numbers now that can sound like bureaucratic gibberish. The U.S. economy grew at an annual rate of 2.4 percent in the last three months of the year, according to government figures released this morning. Got it? Jacob Goldstein from our Planet Money team sheds some light on what that means.

JACOB GOLDSTEIN, BYLINE: If you'd asked somebody 100 years ago, how's the economy doing? Is it growing? Is it shrinking? They would not have known what you were talking about. Back then, people talked about banking panics and national wealth, and trade. But this thing we call the economy wasn't really invented until the 20th century.

ZACHARY KARABELL: It was invented because of the Great Depression.

GOLDSTEIN: Zachary Karabell is the author of a new book called "The Leading Indicators."

KARABELL: And it was invented because there was clearly a perception that there was something really, really bad going on but they didn't really know what. I mean, you could see there were homeless people on the street, you could see there were farmers, you know, the Okies heading from their Dust Bowl farms off to California by the tens of thousands, but there was no way of really grasping it.

GOLDSTEIN: So the government starts calculating this single, official number called national income. It's the forerunner of today's Gross Domestic Product, GDP, and it's basically the value of all the goods and services produced in the country in a year. When it's released in the Depression, this wonky statistic becomes an overnight sensation.

A report on national income submitted to Congress makes the bestseller list. And pretty soon, you can't turn on the radio without hearing those numbers and what they're measuring, this new thing called the economy.

KARABELL: Then, you start hearing about in 1937, Roosevelt starts talking about the economy and he starts talking about national income going up.

(SOUNDBITE OF ARCHIVED BROADCAST)

PRESIDENT FRANKLIN D. ROOSEVELT: That national income had amounted in the year 1929 to $81 billion dollars.

KARABELL: You'd never hear Abraham Lincoln or Teddy Roosevelt or George Washington talking in this way. One of the things that's remarkable to me is how quickly we went from a world where none of these terms and none of this conversation was part of our national consciousness to it being at the center of our national consciousness.
GOLDSTEIN: In the decades that follow, national income becomes gross national product and eventually GDP and it sweeps the world.

KARABELL: The first thing you do in 1950s and '60s if you're a new nation is you open a national airline, you create a national army, and you start measuring GDP.

GOLDSTEIN: That's because if you want help from the World Bank or the U.N., they're going want to know, how does it affect your economy? But somewhere around this time, Karabell says, people start to make too much of GDP. Rather than a limited measure of the economy, it becomes this Cold War gauge of who's doing better or who's winning. And, so, perhaps inevitably, all that success leads to the GDP backlash.

Robert Kennedy famously calls out its shortcomings in 1968.

(SOUNDBITE OF ARCHIVED BROADCAST)

ROBERT KENNEDY: Gross National Product does not allow for the health of our children, the quality of their education, or the joy of their play. It does not include the beauty of our poetry or the strength of our marriages.

DIANE COYLE: It does what says on the tin, it measures the economy. We shouldn't make it do something it was never intended to do.

GOLDSTEIN: Diane Coyle is an economist who just wrote a book called "GDP, A Brief But Affectionate History." Coyle says GDP was never intended to measure overall well being or a nation's standard of living. Certain things that are clearly bad actually make GDP go up, like hurricane damage that costs a lot to fix. And here's another thing. The history of GDP is full of debates about what you should count when you're adding it up.

For example, should you count the black market, which means everything from off-the-books babysitters to mafia drug deals? The U.S. doesn't, other countries do. Back in the '80s, Italy started counting its black market and overnight the Italian economy became bigger than the U.K. economy. The Italians celebrated. They called it Il Sorpasso. Coyle says this points to a common misconception.

COYLE: We tend to think about GDP as if it's a natural object. It's like a mountain, and we have methods of measuring it that are better or worse and more or less accurate. But there is a thing there to be measured. And actually, that's not just true with the economy. There's no natural entity called GDP in the universe.

GOLDSTEIN: In other words, maybe the most important thing to remember about GDP is it's not a thing, it's an idea. And that idea keeps changing. Just last year, the U.S. tweaked the way it calculates GDP and in an instant, the economy was $500 billion bigger. Jacob Goldstein, NPR News.
Fans of the Netflix series "House of Cards" know that part of the show's success is in the details. For a storyline on China, the show consulted political scientists, and some episodes feature actors speaking Chinese. But as NPR's Hansi Lo Wang reports, listening closely, that's one detail the show didn't get quite right.

(SOUNDBITE OF MUSIC)

HANSI LO WANG, BYLINE: Fellow binge-watchers, you'll remember there was a character in the first season of "House of Cards" who spoke in heavily accented Mandarin Chinese.

(SOUNDBITE OF TV SHOW, "HOUSE OF CARDS")

GERALD MCRANEY: (As Raymond Tusk) (Foreign language spoken)

WANG: You wouldn't expect St. Louis billionaire Raymond Tusk to be a fluent speaker. But in the show's second season, there are a few roles that would call for actors to perform in Chinese fluently. So, I turned to an expert.

KIRSTEN SPEIDEL: (Foreign language spoken)

WANG: Kirsten Speidel, my Chinese language instructor in college, who says...

SPEIDEL: Because I'm correcting people's pronunciation daily in class, I'm pretty critical when I hear Chinese in American movies.

WANG: And in television. But she hasn't seen the show yet, so I played her an audio clip of a businessman from China who speaks both English and Mandarin.

(SOUNDBITE OF TV SHOW, "HOUSE OF CARDS")

TERRY CHEN: (As Xander Feng) You don't like your soup.

MICHAEL KELLY: (As Doug Stamper) Not good with chopsticks.

CHEN: (As Xander Feng) (Foreign language spoken)
SPEIDEEL: Not a very good accent here. Could be that he knows some Mandarin, but not very good pronunciation of each word.

WANG: And another clip of a Mandarin translator working with a reporter.

(SOUNDBITE OF TV SHOW, "HOUSE OF CARDS")

MOZHAN MARNO: (As Ayla Sayyad) Say the principal investor has strong ties to the White House.

YAN XI: (As Mandarin translator) (Foreign language spoken)

SPEIDEL: This speaker, I feel, is much more fluid and fluent.

WANG: So an A-plus for this one?

SPEIDEL: Yes. Comparatively speaking, yes.

WANG: And, yes, we are nitpicking but over a show that is obsessed with authenticity.

KENNETH LIN: Obviously, we're always trying to get as close to accurate as we can get.

WANG: "House of Cards" staff writer Kenneth Lin wrote the Mandarin dialogue for the show's Chinese characters.

LIN: Whether or not they sound like natives of Beijing or not is certainly questionable. But, you know, if you go to China, people have a lot of different accents.

WANG: Characters from China in American TV shows and movies are often played by Asian-American actors who are not fluent Chinese speakers.

JANET YANG: The assumption is that nobody will notice or care. As it is, people can't really distinguish between Chinese and Japanese and Korean and Vietnamese and any Asian, so Asians tend to get lumped together.

WANG: Producer Janet Yang has worked for decades on films in both China and Hollywood.

YANG: It's been, for the longest time, catering first to American audiences and then the rest of the world just sort of gobbled up everything that was being made here.

WANG: But today, there's more entertainment that's designed to work in both America and in China, which means more demand for dialect coaches like Doug Honorof. He helps actors pull off the illusion of speaking Chinese fluently.

DOUG HONOROF: And it's not just to sound Chinese. They have to be able to act in Chinese. You have to actually be able to own it so much that you can actually then just perform.
ANDY YU: We're trying to help people escape into this world that scriptwriters created.

WANG: Andy Yu is both a Chinese dialect coach and an actor. He says that Hollywood roles for actors of Asian descent are still mostly limited to immigrant or foreign characters. So, language skills are especially important to get past casting directors.

YU: One of the reasons they hire us is because they expect us to know our language and our culture really well. So we have to deliver.

WANG: Lines delivered even in a slightly off accent can ruin the illusion for audiences in the know. But this is one detail that hasn't stopped season two of "House of Cards" from gaining an audience in China. Since its debut on Sohu - China's Netflix equivalent - it's the most-watched American show. Hansi Lo Wang, NPR News.

Track 7

MELISSA BLOCK, HOST:

In a speech today, President Obama laid out a new vision of the global war on terror. He said that more than a decade after the 9/11 attacks, the threat from terrorism has changed and U.S. policy must change with it.

PRESIDENT BARACK OBAMA: As our fight enters a new phase, America's legitimate claim of self defense cannot be the end of the discussion. To say a military tactic is legal or even effective is not to say it is wise or moral in every instance.

BLOCK: Speaking at the National Defense University, the president pledged to be more transparent about the targeted killing of terrorism suspects overseas and he said he was open to reviewing how drones are used.

NPR's Carrie Johnson is here to talk about the speech and what changes it might mean for national security policy. And, Carrie, what is new about what the president had to say today about drones in particular?

CARRIE JOHNSON, BYLINE: First of all, Melissa, the president - perhaps responding to the infamous filibuster by Kentucky Senator Rand Paul earlier this year - stated once and for all he does not intend to use weaponized drones over American skies and soil.

More importantly, more substantively, he talked about using the same standard for targeting both American citizens overseas and foreign citizens overseas. That includes several criteria, including the fact that these people must, in his view, pose a continuing and eminent threat; that there is no way for the host country in which they're residing or hiding can take action against them short of a drone attack.
And that, finally, the president said he intends to evaluate up front whether any civilian casualties might ensue as a result of these attacks and would only carry out an attack in the situation where virtually no civilian casualties would result.

BLOCK: So this amounts to limiting, narrowing the scope of those drone strikes.

JOHNSON: That's exactly right. Finally, he also talked about having a preference for the Pentagon being in charge of the trigger instead of the CIA.

BLOCK: Carrie, one very controversial issue in the drone strike debate has to do with the killing of American citizens overseas. And we now know, from the White House, that four Americans have been killed in drone strikes since President Obama took office. But only one of them, Anwar al-Awlaki, was specifically targeted. What did the president have to say about that in his speech?

JOHNSON: The president said Anwar al-Awlaki had essentially become a senior operational figure in Al-Qaida in Yemen, and that he was continuously directing attacks against Americans, including the 2009 Christmas Day underwear bombing plot, a plot the following year involving planting bombs on cargo planes. And that Awlaki had been targeted for attack a year before the U.S. actually struck him, and that Congress had been notified in advance.

What the president did not talk about though, was three other U.S. citizens who have been killed in drone strikes since 2009; most controversially, al-Awlaki's 16-year-old son, who was killed a few weeks later in Yemen while he was sitting at an outdoor cafe. A lot of civil liberties experts wonder if that was just a tragic mistake for which the U.S. government is never going to take responsibility.

BLOCK: Carrie, the president also said today that he's going to talk to Congress about how drones are used. Where might that lead?

JOHNSON: So, the president says there are a lot of trade-offs here, Melissa. He says he's open to talking with lawmakers about creating a new unit in the executive branch, or even a special court, to review drone targeting decisions on the front-end. But many judges, many federal judges have expressed some concern about whether that's even constitutional or whether they'd be essentially signing death warrants, and whether that's something judges want to do.

The president said there are good and bad reasons to consider all these things. But he's willing to start a conversation.

BLOCK: OK. And very briefly, Carrie, the president also talked about Guantanamo Bay and he had a little news about detainees at that facility.

JOHNSON: The president says it's within his power, Melissa, to lift a ban on transferring several dozen detainees back to Yemen, where they're from, which could restart the process in some
ways, but members of Congress are already out there saying they don't want detainees to be moved out of GITMO.

BLOCK: And specifically to Yemen, because of fears that they will be released from prison and rejoin the fight?

JOHNSON: That's exactly right.

BLOCK: OK, NPR's Carrie Johnson. Carrie, thank you.

JOHNSON: Thank you.

Track 8

MELISSA BLOCK, HOST:

NPR's Laura Sydell is also at South by Southwest Interactive, where she's experiencing something from another realm altogether - a trip into virtual reality.

And Laura, we're talking about something called Oculus Rift. Before you tell us what that is, why don't you tell us where you are?

LAURA SYDELL, BYLINE: Yes, you will hear theme music from "Game of Thrones" in the background. I am at an HBO exhibition here, where there are costumes from "Game of Thrones," and characters and all kinds of things all around me, from "Game of Thrones."

BLOCK: And "Game of Thrones" is both a both a TV show and a video game. And now, it's also part of this Oculus Rift. So what are we talking about?

SYDELL: All right, yeah. Oculus Rift is a set of goggles. It kind of looks like ski goggles. But what the Oculus Rift can do is make you feel - you don them, and you actually feel as if you are in a completely different world. So you turn your head, and you see the same world.

And I got a chance to try the Oculus Rift because they set up a "Game of Thrones" display here. And what happens is, you get into this box, and it makes it seem like you are going up the winch elevator at Black Castle and that you reach the top, and look out over the mountains.

And I actually could not talk to you from inside this little booth with the Oculus Rift on, 'cause I would have gotten too dizzy. So instead, I'm going to play you tape of me reacting to this experience of donning the Oculus Rift and getting in the booth.

(SOUNDBITE OF COMMENTS TAPE DURING OCULUS RIFT EXPERIENCE)

SYDELL: Wow. I am looking out over a frozen field, and I really feel...

(LAUGHTER)
SYDELL: ...like I am going up an elevator. I see soldiers coming up in the snow, holding torches. Oh, my God.

(LAUGHTER)

SYDELL: And fiery arrows. Ah, I've been hit by a fiery arrow. Oh, my God. Whoa!

BLOCK: Laura, I gather you survived the attack.

SYDELL: I did, and it was hard because I'm afraid of heights.

(LAUGHTER)

BLOCK: Oh, no.

SYDELL: And it was really scary.

BLOCK: Is this interactive, Laura? Is the virtual world responding to things that you do?

SYDELL: You know, in this particular display, it isn't. But I tried a couple of other experiences with the Oculus Rift. In one of them, I was watching an interactive documentary about people who make art with code. And by staring at a particular spot on the screen for a little extra time, I would open up a whole 'nother documentary. So it was responding to my eye movements.

At the same area, I tried something that was not interactive but was amazing. I donned the goggles, and I was suddenly in a musician's studio. And I was sitting right next to the musician as he was playing the keyboards. His dog was on the floor, and I wanted to reach out and pet the dog. It really felt like I could but unfortunately, I can't yet.

I think coming down the road, there are some other things coming up that will make this technology more interactive. But for now, it just really gives you a sense of being there because it responds to a turn of the head and your eye motions.

BLOCK: And is it something that's available for consumers? Or is it really just for the folks at South by Southwest to experience?

SYDELL: Well, right now, developers are creating things like the HBO experience, like that interactive documentary that I mentioned. I am told - I spoke with one of the founder of the company Oculus, Palmer Luckey - and he said we hope to actually get it into a consumers' hands, and more people will be able to have the amazing experiences I just had.

BLOCK: That's NPR's Laura Sydell at South by Southwest Interactive in Austin, Texas. Laura, thanks so much.

SYDELL: You're welcome.
Track 9

JACKI LYDEN, HOST:

OK. So here's a joke. A man sitting on the veranda with his wife one night when out of the blue he says: I love you. His wife says: Was that you or was that the beer talking? The man says: That was me talking to the beer. Maybe you found that funny. I find it hysterical. What makes a joke funny is a question that has beset the human condition since we lost our tails and started walking upright.

But, can scientists tell us today why we laugh? Scott Weems is a cognitive neuroscientist and the author of the book, "Ha! The Science of When We Laugh and Why." And he joins us from member station KUAR in Little Rock, Arkansas. Thanks for being with us, Scott Weems.

SCOTT WEEMS: Thank you very much for having me.

LYDEN: Tell me a joke and let's break it down a little bit, would you?

WEEMS: Sure. I should warn you though, as a scientist I'm not trained very well to tell jokes, but I'll do my best.

LYDEN: All right.

WEEMS: So a dog walks in a telegraph office and he says I want to send a message. And the operator says, "Sure what would you like to send?" And the dog says, "Woof woof woof woof woof woof woof woof woof woof woof." And the operator pauses a second and goes, "You know, that's only nine. You can send a 10th woof for free." And the dog replies, "But that would make no sense." People never laugh when I tell that joke.

LYDEN: I'm giggling. I'm giggling here.

WEEMS: Thank you. I wouldn't hold it against you if you didn't.

LYDEN: All right. I'm kind of getting my comedy writing sketch part here, but you're a neuroscientist. What makes it funny?

WEEMS: There is one part of the brain that's worth recognizing and it's called the anterior cingulate. It's not on the surface. It's a little below and it's what we consider our conflict detector.

LYDEN: Um-hum.

WEEMS: Anytime we're confused or overwhelmed or just we have conflicting information like in the form of a surprising punch line, this area gets very active.

LYDEN: Why is conflict important in a joke?
WEEMS: It's basically how we process things we don't understand. I mean, so much of our life is filled with conflict, and not just jokes. I mean, people laugh at funerals, people laugh at tragic events, and it's because these are time when we just don't know how else to respond. I mean, humor is much broader than just a standup routine. It's just how we look at these moments in life where things don't make sense.

LYDEN: So laughter is coping, bonding, lessening anxiety, the sense of discovery, surprise; all these things.

WEEMS: It is and I think that's why it's so linked with health benefits as well.

LYDEN: Now, humor hasn't always been looked upon so positively. You write that Plato and Jesus weren't funny.

WEEMS: No. I mean, it's really a shame. Historically, humor has not gotten a good rap. Someone actually counted the number of times that laughter occurred in the Old Testament. The total number is 29 and of those only two are positive. In other words, only two are occasions of joy. Well, there's debates now whether Jesus laughed, not just in the New Testament, but in his whole life.

And, of course, Plato, Hobbs, Nietzsche, these scholars all have very negative views towards humor 'cause they saw it as something that weak minds did. It's not something that serious people do.

LYDEN: Do you think men and women tell jokes differently?

WEEMS: They do. It turns out that women laugh more than men, but they're much less successful in the world if comedy. Or at least there are fewer professional female comedians, which is - it's a shame. And people have wondered why is this? Because it's certainly not that women have less of a sense of humor. And one evolutionary theory is that the men are raised, and maybe even have an evolutionary benefit to being the funny people in relationships.

We men make women laugh because it's a sign of genetic fitness. A man who can make his partner laugh is more likely to be intelligent and a good caregiver. And that's also why women consistently rate sense of humor as No. 1 desired trait in a mate. For men, sometimes it's closer to No. 3, after intelligence and good looks. So women maybe aren't given the benefits and the encouragement to be as funny as they could or should be.

LYDEN: You want leave us with one last joke?

WEEMS: Oh, my. OK, I have had bad success with this in the past but I will give it a try. Two fish are swimming in a tank and one looks to the other and he says: Do you know how to drive this thing?
LYDEN: Scott Weems is a cognitive neuroscientist and the author of "Ha! The Science of When We Laugh and Why." He joined us from Little Rock, Arkansas. Thank you for joining us.

WEEMS: Thank you very much for having me.

(SOUNDBITE OF MUSIC)

LYDEN: And you're listening and maybe laughing to WEEKEND EDITION from NPR News.

Track 10

JACKI LYDEN, HOST:
The crisis in Ukraine has many in this country wondering what on earth Vladimir Putin is thinking. Hillary Clinton compared him to Hitler; many world leaders have called his actions insane in recent weeks. How is it that we know so much about Russia's president and yet so little? To help us with that, we've called in someone who's spent a lot of time thinking about Vladimir Putin. Masha Gessen is the author of a best-selling biography of Putin called "The Man Without a Face." Masha Gessen, thank you for joining us.

MASHA GESSEN: Thank you for having me.

LYDEN: So, you wrote an op-ed in the L.A. Times this week with the headline: Is Vladimir Putin Insane? And then you answered your own questions - hardly. So, could you explain, please?

GESSEN: I think Putin has a very consistent worldview. He thinks that what he is doing is right. Everything that he sees around him confirms that what he's doing is right. Among other things, he's really boosted his popularity with the Ukraine effort. His approval ratings dropped around the time that the Russian protest movement erupted and they didn't actually recover until last week.

LYDEN: You are writing that Putin is obsessed with imminent catastrophe and annexing territory. What motivates him?

GESSEN: What motives him is he actually has recently discovered that he has a civilizational mission. He wants a Russia that will become the traditional values capital of the world, that would hold back Western encroachment. This is what he sees happening in Ukraine. He's carrying out his historical mission. A subtext of it is also recreating the Soviet Union and sort of gathering Russian lands, but that's not even the most important part at this point. The most important part is that he thinks that Russia has a unique place in the world and a unique civilization to protect.

LYDEN: Is there a particular political ideology here or would you say rather a conservative moralistic anti-Western ideology that's driving his actions?

GESSEN: That's exactly right. It's a conservative, moralistic anti-Western ideology. It is mostly based on negatives. There's very little that he can say in the affirmative except for the very vague
notion of traditional orthodox culture. But the negatives are actually effective in mobilizing a population, in creating a sense of fear and imminent danger.

LYDEN: Has he been persuasive, do you think, internally, with his arguments that what he's basically doing is protecting Russian interests and people?

GESSEN: Oh, yeah. I mean, the latest opinion polls are absolutely mindboggling. Only 6 percent of Russians, according to a poll by the Levada Center, which is an independent polling organization, only 6 percent of Russians believe that Russia should not be invading Ukraine. The vast majority of Russians believe that Russian speakers and ethnic Russians are in danger in Ukraine. They believe that there is anarchy and no government in Ukraine. And they believe that the invasion is warranted.

LYDEN: Well, the view here in the United States is certainly changing. The whole fundamental disconnect is really quite difficult for people to absorb. Surely, this is affecting U.S.-Russia relations.

GESSEN: Well, I feel a little exasperated. You know, it's about time that the view of him in the United States were changing. I still think there is a lot of misguided conversation that is, again, based on the American worldview. There is conversation about how to help put some save face and pull out of Ukraine. There is no issue of Putin saving face by going out on Ukraine. That's sort of not on the agenda. He is benefiting by being in Ukraine. He's doing exactly what he wants to do and he's getting the results that he wants.

LYDEN: So, what should the U.S. do in this case, do you think?

GESSEN: Measures such as real economic sanctions - not symbolic economic sanctions - but sanctions that would actually affect the Russian elite, Russian business, Russia's place on Western markets. This is all possible. Those would have an impact on the Russian economy. But we have to understand that they will also accelerate the mobilization in Russia and actually lead Putin to escalate the war effort. And if a military response were even on the table, which it's not, but if it were, it would escalate the mobilization even further. So, would any one of the three avenues basically leading to a dead end strategically. Politicians have to ask themselves what's the right thing to do. And I think the right thing to do is to isolate the dictator, to turn Putin into the pariah state that he has put so much effort into creating.

LYDEN: Masha Gessen is the author of "The Man Without a Face: The Unlikely Rise of Vladimir Putin." Masha Gessen, thank you very much.

GESSEN: Thank you for having me.

Track 11

RENEE MONTAGNE, HOST:
Over the coming weeks, NPR will be reporting on women and money - how women save, earn and access it worldwide. It's part of an ongoing focus on how women's lives are changing in this century. One person who is keenly invested in moving women up the economic ladder is Christine Lagarde. She is head of the International Monetary Fund, the IMF. The IMF promotes the stability of the global economy through lending, forecasting and technical assistance. Lagarde has been using her position in the last two and a half years to challenge global leaders to focus specifically on women's economic empowerment as a way to bring about growth generally. She joined us from her office at the IMF in Washington, D.C. Good morning.

CHRISTINE LAGARDE: Good morning.

MONTAGNE: Now, under your leadership, the IMF commissioned a major study on women in the workforce. It finds that more inclusiveness would bring about greater economic growth. And the reverse of that, you might say, is that those countries that don't make use of their potential female workforce, that they're really missing out economically.

LAGARDE: Um-hum. We found some really interesting numbers. If female were working in the same proportion as men do, the level of GDP would be up 27 percent in a country like India, but also up 9 percent in Japan and up 5 percent in the United States of America. It's not just a moral issue, not just a philosophical issue. It just makes economic sense. You know, I was going to say it's a no-brainer.

MONTAGNE: It would seem so when it's laid out like that, but do you imagine that people think of it ever economically or is it mostly always thought of as just a good thing?

LAGARDE: You know, if it was just a good thing to do, it wouldn't go very far. And I'll give you two examples, the example of Japan and the example of Korea. In both those countries the policymakers have decided to put women at the center of their budget, at the center of their policies going forward. For instance, Prime Minister Abe. He's identified a big-budget item that will go to build child care centers in Japan, which is obviously one of the ways to lower the burden on women and facilitate their access to the job market.

MONTAGNE: Well, part of the issue here about bringing women to the workforce is not so much just bringing them to the workforce because, as your study notes, women are in lots of jobs that are low paid and have very little power. When you talk about management, one thing that's been noted since you were named the head of the IMF is that it itself has a stunningly low number of women in management, high management positions - only about 20 percent there.

LAGARDE: Correct.

MONTAGNE: What made it so hard for the IMF to get women in these higher positions?
LAGARDE: One of the reasons has to do with the population from which management are drawn. Most of them are economists. All of them in those positions are generally Ph.Ds. And when you look for the population of female Ph.Ds. from which you can draw those management skills, there are not that many of them. So, when you start with a relatively small pool, it's probably a little bit harder and it requires much more of an effort. But we do have targets. And we are certainly going to continue to be focused on those.

MONTAGNE: Well, you know, something that just struck me when you said that. In a way, we could be saying that there are just not enough women who have the qualifications for being in management. If a man said that, would he get in trouble?

LAGARDE: If a man said that to me, I think he might get in trouble, yes. There are now more and more young, talented female economists. So, if any head of department tells me, no, I'm terribly sorry, I can't hire a woman because I can't find any talented or competent women, I would say rubbish. In the 2013 Economists Program, we hired 51 percent women, 49 percent men. And the reason for that is that we have a draft from all over the world and we've hired, for instance, in that group a good number of Chinese economists - highly qualified, all Ph.Ds. from the best universities of the world. And guess what? They're all women.

MONTAGNE: Well, you yourself must have a great deal of experience with being the only woman in the room. And I gather that was even true when you were being interviewed to take over at the IMF.

LAGARDE: Well, yeah, that is true because the IMF has an executive board, which includes 24 members. And of the 24 people in the room, there were 24 men because the only female executive director was actually away on the very technical work that we do. And I can't do anything about it except talk about it and complain about it when I talk about it because those people are actually appointed by the various countries around the world that comprise the membership of the IMF.

MONTAGNE: You are so confident, that's quite clear. So, did you not run into discrimination...

LAGARDE: Oh, yes, yes, yes, of course. Yeah. When I started my life as a baby lawyer, I was interviewing with all the best firms then at the time, and one of them was, you know, probably the most reputable in France, and told me you can join us tomorrow and be an associate, be given great tasks and great files and great clients to work on, but don't ever expect to make partnership. And I said, you know, why would that be? And they said because you're a woman. And I said really? Well, you won't have me as an associate. And I went to another firm which was great and which was based on respect and on equal opportunity for those who delivered.

MONTAGNE: You were quoted once as saying that the financial mess would not have been such a mess if it had been Lehman Sisters instead of Lehman Brothers. What exactly did you mean? And does that apply - what you were saying apply, generally?
LAGARDE: I would repeat that. And I do believe women have different ways of taking risks, of ruminating a bit more before they jump to conclusions. And I think that as a result, particularly on the, on, you know, on the trading floor, in the financial markets in general, the approach would be different. You know, I'm not suggesting that all key functions and roles should be held by women. But if you look at the studies - and there were quite a few that were done by the Financial Times, by the Economist, by various financial observers of the market - it's apparently very clear now that those companies that have several female directors on their board and female in their top management actually do better, are more profitable, and give a better return to their shareholders.

MONTAGNE: Christine Lagarde, managing director of the International Monetary Fund. Thank you so much for joining us.

LAGARDE: Thank you.

(SOUNDBITE OF MUSIC)

STEVE INSKEEP, HOST:

Hey, we asked Christine Lagarde if she'd share a money lesson from her own life, and you can see her answer, and provide a lesson you've learned as well, on our Tumblr. Just search for She Works Tumblr. You'll find it. It's worth the search. Also, thoughts from Olympia Snowe, Neko Case and more. She Works Tumblr. This is NPR News.

Track 12

LINDA WERTHEIMER, HOST:

Students sizing up law schools may take into account how many graduates from a particular institution actually get employed once they have their degrees in hand. Many schools promise jobs after graduation, but not all those legal jobs are quite what they seem.

Ashley Milne-Tyte from our Planet Money team has this report.

ASHLEY MILNE-TYTE, BYLINE: Prospective law students who can't make it to Williamsburg, Virginia can learn all about the law school there on the Web.

(SOUNDBITE OF MUSIC)

UNIDENTIFIED MAN: William and Mary is the oldest law school in the United States.

MILNE-TYTE: Complete with elegant buildings, famous names...

(SOUNDBITE OF MUSIC)

UNIDENTIFIED MAN: Founded in 1779 at the urging of Thomas Jefferson.
MILNE-TYTE: And if you delve into their website, you can see the graduates of William and Mary Law School are doing great in the job market - they have a 90 percent employment rate. But not all those positions are big-paying law firm jobs. When you dig down, there's a surprising fact: A fifth of graduates are employed by the university itself. They're not working for the school. They're in a fellowship program that pays graduates a stipend to work in public service jobs or for non-profits. Dave Douglas is dean of the law school.

DAVE DOUGLAS: In this market, where jobs are tight, a student needs to have the opportunity to show what they can do.

MILNE-TYTE: He says with these jobs, graduates gain valuable experience. Within a year, most of the university-funded graduates land full-time jobs.

DOUGLAS: And that's what this fellowship program does, and that's why these students succeed.

MILNE-TYTE: It does something else too - it boosts the school's place in the Best Law Schools rankings produced by U.S. News & World Report. William and Mary Law School jumped nine spots this year. It's now the 24th best school in the nation. The dean says that jump happened in part because of the school's improved employment numbers.

Moving up in the rankings is a big deal.

KAREN SLOAN: I think it has an outsized influence, this sort of rankings obsession, in the legal profession, which is very sort of obsessed with prestige.

MILNE-TYTE: Karen Sloan is a reporter with the National Law Journal. She says a lot of things go into these rankings, but one crucial factor is how many graduates land jobs. The rankings count someone employed if they have a professional job nine months after graduation, but they don't care who employs you. So, many top-ranked law schools have these school-funded job programs, Georgetown, NYU, George Washington University.

Sloan says these kinds of initiatives became popular during the recession when legal jobs got scarce.

SLOAN: So there's really no doubt in my mind that the primary motivation for these programs was to boost employment numbers.

MILNE-TYTE: Now there's been debate in the past about how transparent law schools have been about their job numbers. Until fairly recently, you could be working at McDonald's after graduation and you'd still count as employed in the rankings. Then the rules changed. Now schools provide a more detailed breakdown of the kinds of jobs graduates get.
But Kyle McEntee of the non-profit Law School Transparency, he says there are still problems with these school-funded programs.

**KYLE MCENTEE:** They're not purely intentioned. They exist to help appearances whether it's improving the rankings criteria for U.S. News rankings, or improving their employment rates.

**MILNE-TYTE:** McEntee says students can't make an informed choice about the return on their investment if they don't know which are real jobs, and which aren't. Law schools say look, anyone can see the nitty gritty of our employment numbers on our websites - we make it clear how many graduates are on our dime.

And the law school students we spoke to didn't feel like anything nefarious was going on. Andrew Beyda goes to George Washington University Law School. He says any job, even one funded by your school, is better than no job.

**ANDREW BEYDA:** Well, I mean it's a tough legal market, I mean especially since the economic crisis around 2008, frankly lawyers aren't retiring or dying nearly fast enough in order for us to fill their spots.

**MILNE-TYTE:** And the schools that run these programs make another point: These temporary jobs are a way to get students into public service. Brian Daner graduated from the University of Virginia Law School in 2011. He had a job offer from a private law firm in Washington, D.C. But he turned it down to go and work in a temporary position on Capitol Hill. A University of Virginia fellowship made it possible for him to do that.

**BRIAN DANER:** And the fellowship program was actually a year long, but after five months, I guess I proved my worth to my bosses and they brought me on as full-time counsel.

**MILNE-TYTE:** Some of his classmates also graduated to jobs in public service right from their fellowships.

Karen Sloan of the National Law Journal agrees the programs are a great opportunity for students at the universities that offer them. But she says overall, a small percentage of the country's law school graduates even have this option.

**SLOAN:** Some of the schools lower down the U.S. News the rankings, just, I mean they don't have the kind of endowment money, they don't have the finances to do something like this.

**MILNE-TYTE:** If you're one of those schools that doesn't have an employment program, your students are on their own in a legal job market that still hasn't recovered from the recession.

Ashley Milne-Tyte, NPR News.

**WERTHEIMER:** And that's the business news on MORNING EDITION from NPR News. I'm Linda Wertheimer.
STEVE INSKEEP, HOST:

And I'm Steve Inskeep.
Appendix B:  
Target Word Lists used in SAS

The following are the target word lists used for both SAS and SAS_v2 tasks. The first number is the time of word on-set relative to the start of the track, and the latter number is the time for end of word presentation. All time values are in milliseconds.

Track 1

<table>
<thead>
<tr>
<th>Word</th>
<th>On-Set Time</th>
<th>End-Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>speaker</td>
<td>5070</td>
<td>5390</td>
</tr>
<tr>
<td>clear</td>
<td>15310</td>
<td>16160</td>
</tr>
<tr>
<td>common</td>
<td>19900</td>
<td>20250</td>
</tr>
<tr>
<td>house</td>
<td>24440</td>
<td>24640</td>
</tr>
<tr>
<td>president</td>
<td>31340</td>
<td>31740</td>
</tr>
<tr>
<td>months</td>
<td>34780</td>
<td>35270</td>
</tr>
<tr>
<td>why</td>
<td>39690</td>
<td>40010</td>
</tr>
<tr>
<td>fiscal</td>
<td>45070</td>
<td>45590</td>
</tr>
<tr>
<td>looking</td>
<td>49890</td>
<td>50160</td>
</tr>
<tr>
<td>weeks</td>
<td>54050</td>
<td>54240</td>
</tr>
<tr>
<td>up</td>
<td>63580</td>
<td>63850</td>
</tr>
<tr>
<td>engage</td>
<td>66280</td>
<td>67220</td>
</tr>
<tr>
<td>american</td>
<td>72210</td>
<td>72540</td>
</tr>
<tr>
<td>him</td>
<td>77220</td>
<td>77580</td>
</tr>
<tr>
<td>first</td>
<td>82750</td>
<td>82980</td>
</tr>
<tr>
<td>hints</td>
<td>87250</td>
<td>87600</td>
</tr>
<tr>
<td>bigger</td>
<td>93690</td>
<td>93960</td>
</tr>
<tr>
<td>shutdown</td>
<td>96580</td>
<td>97330</td>
</tr>
<tr>
<td>well</td>
<td>105120</td>
<td>105320</td>
</tr>
<tr>
<td>talked</td>
<td>109740</td>
<td>110000</td>
</tr>
<tr>
<td>law</td>
<td>114370</td>
<td>114630</td>
</tr>
<tr>
<td>room</td>
<td>122860</td>
<td>123420</td>
</tr>
<tr>
<td>stalled</td>
<td>126200</td>
<td>126800</td>
</tr>
<tr>
<td>both</td>
<td>130930</td>
<td>131170</td>
</tr>
<tr>
<td>plan</td>
<td>138230</td>
<td>139220</td>
</tr>
<tr>
<td>oval</td>
<td>144120</td>
<td>144400</td>
</tr>
<tr>
<td>drive</td>
<td>148610</td>
<td>148890</td>
</tr>
<tr>
<td>broken</td>
<td>154930</td>
<td>155600</td>
</tr>
<tr>
<td>abuse</td>
<td>160960</td>
<td>161730</td>
</tr>
<tr>
<td>again</td>
<td>168620</td>
<td>169780</td>
</tr>
</tbody>
</table>
budget 175550 175850
include 180310 180790
right 186990 187150
lots 193960 194170
reduced 199010 199680
idea 204080 204720
them 208020 208230
coupled 213050 213400
revenue 220080 220490
nowhere 224020 224750

Track 3

doctors 5080 5710
new 10410 10580
decline 14910 15420
deadline 23640 24100
us 30480 30620
first 36040 36340
think 45780 46010
health 53400 53580
earn 60730 60890
sales 67700 68050
cash 74550 75000
years 84160 84440
cases 90640 91520
exchange 98220 98680
medical 107840 108180
drug 112720 112930
slack 117570 118040
cut 123980 124230
kind 129370 129600
huge 134700 134980
sales 140280 140600
big 148690 148880
tend 156590 156940
wave 164430 164730
money 168960 169250
are 174330 174420
shining 184750 185340
<table>
<thead>
<tr>
<th>Term</th>
<th>191730</th>
<th>191980</th>
</tr>
</thead>
<tbody>
<tr>
<td>health</td>
<td>202610</td>
<td>202850</td>
</tr>
<tr>
<td>all</td>
<td>207330</td>
<td>208090</td>
</tr>
<tr>
<td>industry</td>
<td>213120</td>
<td>213260</td>
</tr>
<tr>
<td>that</td>
<td>222680</td>
<td>222890</td>
</tr>
<tr>
<td>health</td>
<td>228680</td>
<td>229220</td>
</tr>
<tr>
<td>commerce</td>
<td>233640</td>
<td>233820</td>
</tr>
<tr>
<td>do</td>
<td>242640</td>
<td>242940</td>
</tr>
<tr>
<td>next</td>
<td>249860</td>
<td>250050</td>
</tr>
<tr>
<td>other</td>
<td>257610</td>
<td>258130</td>
</tr>
<tr>
<td>sunshine</td>
<td>265820</td>
<td>266100</td>
</tr>
<tr>
<td>folks</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Track 5**

<table>
<thead>
<tr>
<th>Term</th>
<th>5030</th>
<th>5530</th>
</tr>
</thead>
<tbody>
<tr>
<td>economy</td>
<td>10210</td>
<td>10640</td>
</tr>
<tr>
<td>government</td>
<td>14680</td>
<td>14890</td>
</tr>
<tr>
<td>money</td>
<td>20330</td>
<td>20790</td>
</tr>
<tr>
<td>doing</td>
<td>29780</td>
<td>29890</td>
</tr>
<tr>
<td>we</td>
<td>34870</td>
<td>35220</td>
</tr>
<tr>
<td>because</td>
<td>39770</td>
<td>39820</td>
</tr>
<tr>
<td>it</td>
<td>45910</td>
<td>46120</td>
</tr>
<tr>
<td>know</td>
<td>55430</td>
<td>55600</td>
</tr>
<tr>
<td>there</td>
<td>60260</td>
<td>60730</td>
</tr>
<tr>
<td>number</td>
<td>64850</td>
<td>65440</td>
</tr>
<tr>
<td>product</td>
<td>70060</td>
<td>70530</td>
</tr>
<tr>
<td>country</td>
<td>80000</td>
<td>80470</td>
</tr>
<tr>
<td>list</td>
<td>85620</td>
<td>85870</td>
</tr>
<tr>
<td>thing</td>
<td>91720</td>
<td>92010</td>
</tr>
<tr>
<td>starts</td>
<td>100260</td>
<td>100450</td>
</tr>
<tr>
<td>in</td>
<td>106530</td>
<td>107040</td>
</tr>
<tr>
<td>lincoln</td>
<td>114960</td>
<td>115420</td>
</tr>
<tr>
<td>terms</td>
<td>120840</td>
<td>121010</td>
</tr>
<tr>
<td>at</td>
<td>125050</td>
<td>125410</td>
</tr>
<tr>
<td>income</td>
<td>130490</td>
<td>130810</td>
</tr>
<tr>
<td>first</td>
<td>135810</td>
<td>136200</td>
</tr>
<tr>
<td>create</td>
<td>145290</td>
<td>145640</td>
</tr>
<tr>
<td>time</td>
<td>150250</td>
<td>150680</td>
</tr>
<tr>
<td>measure</td>
<td>154790</td>
<td>155580</td>
</tr>
<tr>
<td>winning</td>
<td>162070</td>
<td>162400</td>
</tr>
<tr>
<td>calls</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

87
<table>
<thead>
<tr>
<th>Word</th>
<th>Track 1</th>
<th>Track 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>do</td>
<td>180390</td>
<td>180590</td>
</tr>
<tr>
<td>brief</td>
<td>185950</td>
<td>186240</td>
</tr>
<tr>
<td>being</td>
<td>190700</td>
<td>191160</td>
</tr>
<tr>
<td>bad</td>
<td>194490</td>
<td>195080</td>
</tr>
<tr>
<td>another</td>
<td>200100</td>
<td>200390</td>
</tr>
<tr>
<td>example</td>
<td>206010</td>
<td>206610</td>
</tr>
<tr>
<td>mafia</td>
<td>211230</td>
<td>211720</td>
</tr>
<tr>
<td>italy</td>
<td>216310</td>
<td>216640</td>
</tr>
<tr>
<td>bigger</td>
<td>220800</td>
<td>221110</td>
</tr>
<tr>
<td>says</td>
<td>226010</td>
<td>226400</td>
</tr>
<tr>
<td>object</td>
<td>231350</td>
<td>232470</td>
</tr>
<tr>
<td>measuring</td>
<td>235080</td>
<td>235520</td>
</tr>
<tr>
<td>be</td>
<td>240290</td>
<td>240390</td>
</tr>
<tr>
<td>entity</td>
<td>246190</td>
<td>246710</td>
</tr>
<tr>
<td>maybe</td>
<td>250320</td>
<td>250580</td>
</tr>
<tr>
<td>idea</td>
<td>256480</td>
<td>256780</td>
</tr>
<tr>
<td>instant</td>
<td>262470</td>
<td>263210</td>
</tr>
</tbody>
</table>

**Track 7**

<table>
<thead>
<tr>
<th>Word</th>
<th>Track 1</th>
<th>Track 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>terror</td>
<td>4960</td>
<td>5920</td>
</tr>
<tr>
<td>changed</td>
<td>10490</td>
<td>11370</td>
</tr>
<tr>
<td>our</td>
<td>14060</td>
<td>14620</td>
</tr>
<tr>
<td>end</td>
<td>20680</td>
<td>20830</td>
</tr>
<tr>
<td>say</td>
<td>29170</td>
<td>29610</td>
</tr>
<tr>
<td>defense</td>
<td>34260</td>
<td>34700</td>
</tr>
<tr>
<td>overseas</td>
<td>40110</td>
<td>41190</td>
</tr>
<tr>
<td>here</td>
<td>45900</td>
<td>46100</td>
</tr>
<tr>
<td>policy</td>
<td>49240</td>
<td>49710</td>
</tr>
<tr>
<td>melissa</td>
<td>54770</td>
<td>55260</td>
</tr>
<tr>
<td>senator</td>
<td>59410</td>
<td>59870</td>
</tr>
<tr>
<td>over</td>
<td>65820</td>
<td>66210</td>
</tr>
<tr>
<td>more</td>
<td>70100</td>
<td>70310</td>
</tr>
<tr>
<td>citizens</td>
<td>75200</td>
<td>75800</td>
</tr>
<tr>
<td>criteria</td>
<td>80150</td>
<td>81100</td>
</tr>
<tr>
<td>threat</td>
<td>85440</td>
<td>86320</td>
</tr>
<tr>
<td>can</td>
<td>89980</td>
<td>90130</td>
</tr>
<tr>
<td>finally</td>
<td>94680</td>
<td>95020</td>
</tr>
<tr>
<td>casualties</td>
<td>100230</td>
<td>100870</td>
</tr>
<tr>
<td>situation</td>
<td>105240</td>
<td>105820</td>
</tr>
</tbody>
</table>
limiting  110260  110730
talked    114990  115280
one       121130  121380
and       126880  127020
since     132530  132720
say       138660  138830
senior    143090  143590
against   149190  149580
plot      154190  155070
planes    159160  160070
struck    165350  165670
about     170600  170870
since     175040  175340
weeks     181480  181710
civil     185150  185410
never     190680  190930
congress  195150  196140
open      202190  202470
branch    206030  206770
judges    213720  214200
bad       224910  225120
guantanamo 231470  232010
within    236390  236770
several   240390  241250
process   245860  246390
don't     250790  250990
fears     254750  254960

Track 9

love     6480     6820
man      11410    11650
funny    15490    15780
beset    20340    20960
tell     26220    26500
author   31720    32060
member   36650    36890
for      40460    40530
bit      45300    45470
jokes    50220    50520
Track 11

women 3710 4060
focus 9230 9630
ladder 16630 17120
fund 20650 21330
stability 23690 24300
has 29240 29420
challenge 32410 32960
growth 38540 38950
in 42740 42880
now 45820 45980
more 52690 52860
reverse 55890 56320
we 66560 66690
working 70510 71250
level 74910 75160
percent 77140 77970
nine 80930 81220
united 83960 84390
moral 87500 87870
going 93470 93630
you 98270 98360
mostly 102350 102640
korea 115740 117220
put 123250 124240
identified 133230 134100
child 136500 136850
obviously 139000 139460
women 142640 143000
issue 146530 146770
are 153830 154090
very 157510 157690
been 163230 163340
it 167260 167410
high 171040 171340
hard 176400 177510

author 256420 256730
little 260770 260970
<table>
<thead>
<tr>
<th>Word</th>
<th>182220</th>
<th>182510</th>
</tr>
</thead>
<tbody>
<tr>
<td>population</td>
<td>184240</td>
<td>185320</td>
</tr>
<tr>
<td>economists</td>
<td>189600</td>
<td>190950</td>
</tr>
<tr>
<td>generally</td>
<td>192640</td>
<td>193090</td>
</tr>
<tr>
<td>from</td>
<td>198680</td>
<td>198950</td>
</tr>
<tr>
<td>many</td>
<td>202170</td>
<td>202390</td>
</tr>
<tr>
<td>probably</td>
<td>208490</td>
<td>208950</td>
</tr>
<tr>
<td>of</td>
<td>211210</td>
<td>211310</td>
</tr>
<tr>
<td>focused</td>
<td>215920</td>
<td>216350</td>
</tr>
<tr>
<td>way</td>
<td>221260</td>
<td>221640</td>
</tr>
<tr>
<td>he</td>
<td>229300</td>
<td>229430</td>
</tr>
<tr>
<td>i</td>
<td>231950</td>
<td>232040</td>
</tr>
<tr>
<td>so</td>
<td>240720</td>
<td>241450</td>
</tr>
<tr>
<td>because</td>
<td>245060</td>
<td>245370</td>
</tr>
<tr>
<td>rubbish</td>
<td>248510</td>
<td>249430</td>
</tr>
<tr>
<td>hired</td>
<td>252910</td>
<td>253470</td>
</tr>
<tr>
<td>reason</td>
<td>257500</td>
<td>257860</td>
</tr>
<tr>
<td>all</td>
<td>260740</td>
<td>260990</td>
</tr>
<tr>
<td>instance</td>
<td>263100</td>
<td>263540</td>
</tr>
<tr>
<td>qualified</td>
<td>268620</td>
<td>269420</td>
</tr>
<tr>
<td>guess</td>
<td>272560</td>
<td>272770</td>
</tr>
<tr>
<td>deal</td>
<td>278330</td>
<td>278640</td>
</tr>
<tr>
<td>room</td>
<td>281910</td>
<td>283250</td>
</tr>
<tr>
<td>interviewed</td>
<td>285820</td>
<td>286640</td>
</tr>
<tr>
<td>true</td>
<td>291280</td>
<td>291470</td>
</tr>
<tr>
<td>board</td>
<td>293710</td>
<td>294580</td>
</tr>
</tbody>
</table>