NTIA Technical Report TR-09-459

Relationships Between Intelligibility, Speaker Identification, and the Detection of Dramatized Urgency

Andrew A. Catellier
Stephen D. Voran

report series

U.S. DEPARTMENT OF COMMERCE • National Telecommunications and Information Administration

NTIA Technical Report TR-09-459

Relationships Between Intelligibility, Speaker Identification, and the Detection of Dramatized Urgency

Andrew A. Catellier
Stephen D. Voran

U.S. DEPARTMENT OF COMMERCE
Carlos M. Gutierrez, Secretary

Meredith A. Baker, Acting Assistant Secretary
for Communications and Information

November 2008

DISCLAIMER

Certain commercial equipment and materials are identified in this report to specify adequately the technical aspects of the reported results. In no case does such identification imply recommendations or endorsement by the National Telecommunications and Information Administration, nor does it imply that the material or equipment identified is the best available for this purpose.

CONTENTS

Page

FIGURES

TABLES

ABBREVIATIONS/ACRONYMS

DOC	Department of Commerce
DU	Dramatized Urgency
IP	Internet Protocol
ITS	Institute for Telecommunication Sciences
LAN	Local Area Network
LDC	Linguistic Data Consortium
MNRU	Modulated Noise Reference Unit
NTIA	National Telecommunications and Information Administration
NTP	Normalized Task Performance
PCM	Pulse Code Modulation
PDA	Personal Digital Assistant
SID	Speaker Identification
SNR	Signal-to-Noise Ratio
SUSAS	Speech Under Simulated and Actual Stress
TIS	Task Induced Stress
TSID	Tactical Speaker Identification Database

RELATIONSHIPS BETWEEN INTELLIGIBILITY, SPEAKER IDENTIFICATION, AND THE DETECTION OF DRAMATIZED URGENCY

Andrew A. Catellier and Stephen D. Voran[1]

The systems used for public safety speech communications must be intelligible. It is also desirable that they transmit secondary information, such as the attributes of a speaker's voice. This secondary information can allow a user to identify the speaker and his or her emotional state. Testing speech communications systems for the delivery of intelligible speech is common. Testing for human perception of the delivery of this secondary information is less common, though some prior work has been done. Building on this prior work, we describe a set of controlled laboratory listening experiments. These experiments characterize the relationships between speech intelligibility, speaker identification, and the detection of dramatized urgency in a speaker's voice across a range of simulated speech processing conditions. The experiment results indicate that for the speech processing conditions considered here, detection of dramatized urgency is the most robust property, speaker identification is less robust, and speech intelligibility is the least robust.

Key words: human listening tests; intelligible speech; speaker identification; speaker stress detection; speaker urgency detection; speech transmission system; subjective speech quality tests

1 INTRODUCTION

Public safety speech communication systems are designed to carry a message from a speaking user (speaker) to a listening user (listener). The most important characteristic of these systems is intelligibility (i.e., whether the listener can understand the message). In addition to intelligibility, public safety communication systems should aim to successfully transmit secondary information, such as attributes of the speaker's voice. Secondary information allows the listener to identify the speaker or to recognize the emotional state of the speaker.

The ability to identify or confirm the identity of a speaker is called "speaker identification" (SID). SID can be particularly important to public safety officials who rapidly communicate with each other to accomplish time-critical emergency operations. If speakers can be identified implicitly based on transmitted attributes of their voices, the additional overhead associated with explicit identification can be avoided (e.g., "This is Officer Roberts speaking."). Even more valuable is the possibility of detecting whether a speaker has lied about their identity.

Similarly, public safety officials often need to monitor numerous transmissions with only partial attention while simultaneously performing other important tasks. If one of many different

[1] The authors are with the Institute for Telecommunication Sciences, National Telecommunications and Information Administration, U.S. Department of Commerce, Boulder, CO 80305.

speakers associated with one of many different transmissions displays a shift in emotional state via his or her voice, detecting that shift can be very important. When such a shift is detected (e.g., from a neutral tone to a tone of urgency), the listening public safety officials will commit full attention to that specific speaking official to provide support and aid.

This report describes the design, implementation, analysis, and results of a set of three controlled laboratory listening experiments. Each experiment uses a set of six different processing conditions (e.g., speech coders) that spanned a broad range of perceptual quality. The first experiment characterized the relative intelligibility in a sentence context of each processing condition. In this document, if intelligibility is mentioned, it can be assumed that we mean "intelligibility in a sentence context." In the intelligibility experiment, the subjects listened to a sentence and attempted to repeat it verbatim. The second experiment characterized human SID performance for each processing condition. In this SID experiment, listeners learned to recognize six different speakers, and then attempted to recognize each speaker's processed voice. The third experiment characterized a listener's ability to detect dramatized urgency in processed speech. In this experiment, listeners attempted to detect one of two dramatized emotional states based on the voice characteristics of a speaker. These emotional states are "dramatized neutral" and "dramatized urgency" (DU), so we refer to this experiment as "detection of DU." Together, these three experiments characterized the relationships between speech intelligibility, SID, and the detection of DU across six speech processing conditions.

In the sections that follow, we describe related work previously conducted by other researchers. We then describe the various speech recordings used in the three experiments, the six speech processing conditions, the three experiment designs, the software used, and the main results obtained. The results show how intelligibility, SID, and detection of DU vary as a function of the six speech processing conditions used.

2 PREVIOUS RESEARCH

Much work has been done to develop means for testing speech communications systems. However, testing for human perception of the delivery of the secondary information is less common. We are not aware of any previous efforts to characterize how the human detection of speaker emotional state is influenced by the distortions caused by speech processing associated with communication systems. Prior work has been done on the related topics of automatic recognition of speaker emotions [1] and automatic speech recognition that is invariant to speaker emotions [2].

Various studies related to SID have been conducted over the years. In 1963 Compton studied human SID abilities for multiple filtered versions of the sustained vowel sound at the end of the word “three” [3]. He found that SID can happen with recordings as short as 1/40 of a second. He also found that when the pitch of different speakers was closer, those speakers were more easily confused.

Bricker and Pruzansky conducted an experiment where coworkers were asked to identify speakers using processed speech recordings. The speakers were familiar to the listeners, and pictures were used to aid the identification process [4].

Uzdy used two different low-rate vocoders to conduct a SID experiment where listeners were familiar with the speakers [5]. His goal was to determine each vocoder’s effectiveness in transmitting data pertinent to SID. This goal is similar to our current work. Uzdy discussed the importance of adequate listener training and noted that about five hours of training were necessary to obtain stable results.

Schmidt-Nielsen did significant, sustained work on human and machine SID performance, SID performance for familiar and new speakers, and the relations between SID performance and speech coding distortions [6-10]. In [6] she suggests using a small number of speakers to keep within the restrictions of listener memory. Quatieri describes significant work relating machine SID to coding distortions in [11].

3 SPEECH RECORDINGS

This section describes the speech recordings used for each experiment. Example speech files from the DU portion of this experiment are available at <http://www.its.bldrdoc.gov/pub/audio/pubs_talks.php>.

3.1 Intelligibility

There are numerous approaches to testing the intelligibility of speech. Some approaches require the identification of isolated unrelated words, or leading consonants of rhyming words. In public safety communications, as in much of naturally occurring spoken communications, groups of words provide context for each other and that can help a listener to properly understand some hard-to-hear words. We selected "word intelligibility in a sentence context" as a suitable way to assess intelligibility in the present application. To test word intelligibility in sentence context, we selected and recorded 20 sentences from current issues of *The Wall Street Journal* and *The New York Times*. Sentence lengths ranged from 6 to 14 words with a median length of 9 words (e.g., "This rebellion has forced banks to reduce bond offerings."). The sentences were selected to be of average complexity and to contain only commonly used words. The sentences are considered typical in terms of the amount of context the words within a sentence provide for each other. One female and one male speaker recorded each of the sentences. We used studio-grade digital recording equipment and a quiet recording room with average noise level below 20 dBA.

3.2 Speaker Identification

A search for North American English recordings to use in the SID experiment resulted in the selection of the Tactical Speaker Identification Database (TSID), which is available from the Linguistic Data Consortium (LDC) [14]. We chose this database because it includes semi-spontaneous speech, repeated utterances of lists of sentences and digits, and some utterances that are recorded by multiple speakers.

To ensure that the experiment size was manageable within the limits of human memory (as suggested in [8]), we decided to select three female speakers and three male speakers from the database. Average pitch and voicing strength were determined for each male and female speaker. We then looked for male speakers that spoke the same sentences and spanned the full range of pitches found in males in the database. Additional considerations in selecting male speakers and recordings included minimizing speaker script-reading errors, minimizing microphone handling and breath noises, and minimizing microphone overload distortion. We selected three of the four female speakers found in the database by maximizing the range of pitches and the quality of recordings.

After speaker selection, we looked for similar digit sequences (of lengths two and four) and sentences with similar content spoken by each speaker. These were used to form clips of three lengths: short, medium, and long. Semi-spontaneous speech was used for training purposes.

3.3 Detection of Speaker "Stress"

In general, we are interested in listener detection of speaker "stress." But the term "stress" is subjective, and covers a wide range of circumstances and resulting speech signals. For speech signals, objective refinement of the term "stress" and quantification of stressor levels is enabled through the use of known acoustic correlates. These include changes in level, tempo, pitch, and formants [1, 2, 13, 14].

3.3.1 Task Induced Stress

Previous work to develop and test automatic speech recognition that is invariant to speaker emotions resulted in the Speech under Simulated and Actual Stress (SUSAS) recorded speech database [13, 15]. One portion of the SUSAS database involved a male helicopter pilot recording isolated words in neutral (helicopter on the ground and running) and task (pilot flying helicopter) situations. Another portion included one male and one female speaker recording isolated words in neutral (no task) and computer-graphics based "dual tracking" task situations. When comparing the task recordings with the neutral recordings, only a minor sense of distraction is evident. We call this Task Induced Stress (TIS), and we used some of these recordings as the basis for a portion of this experiment. The experiment revealed that the detection of TIS is a very difficult task, independent of the speech processing condition. Because of this, the TIS portion of the experiment is not addressed in the remainder of this report.

3.3.2 Dramatized Urgency

Since urgent speech is not uncommon in public safety communications, speakers conveying urgency were an important part of this experiment. It would not be ethical to subject speakers to events (e.g., physical dangers) that could create a true sense of urgency. Recording speakers confronted by naturally occurring urgency-inducing events was not a practical option, but this might be considered for potential future work. We elected to create recordings of DU.

We monitored public safety communication channels and transcribed messages between public safety personnel to use as scripts. Messages selected ranged in length from 2 words to 21 words with a median length of 9 words (e.g., "We have two children still trapped under the bus."). For comparison purposes, the scripts also included the isolated words of the TIS recordings.

One female and one male speaker recorded the DU scripts. We used studio-grade digital recording equipment and a quiet recording room with average noise level below 20 dBA. Each speaker read the scripts while verbally dramatizing two different situations: a non-urgent situation (neutral) and a situation requiring an urgent response (DU situation). We activated a set of rotating mirrored red and blue strobe lights to provide an unmistakable visual indication of when the speakers should dramatize urgency. A total of 16 different messages were used in the experiment.

3.3.3 Acoustic Correlates in Dramatized Urgency

We analyzed the DU recordings and identified several acoustic correlates. The level of DU speech increased (relative to neutral speech) by an average of 6.2 dB for the male speaker, and 8.0 dB for the female speaker. However, note that this level increase was not directly available to listeners because it was removed via level normalization. It may have been indirectly available if it was accompanied by audible sounds of increased speaking effort.

The two speakers responded oppositely in terms of tempo. The male speaker increased his talking tempo slightly in DU, so his average message duration decreased from 2.86 to 2.68 seconds. The female lengthened certain words for emphasis and thus decreased her tempo. Her average message duration increased from 2.73 to 3.01 seconds.

The mean pitch of the male speaker increased from 134 Hz (neutral) to 148 Hz (DU), while the standard deviation increased from 21 to 23 Hz. For the female speaker, the mean pitch increased from 211 to 249 Hz, and standard deviation increased from 18 to 38 Hz. The pitch histograms in Figure 1 show all four of these results. We also observed changes in formant structure for both speakers.

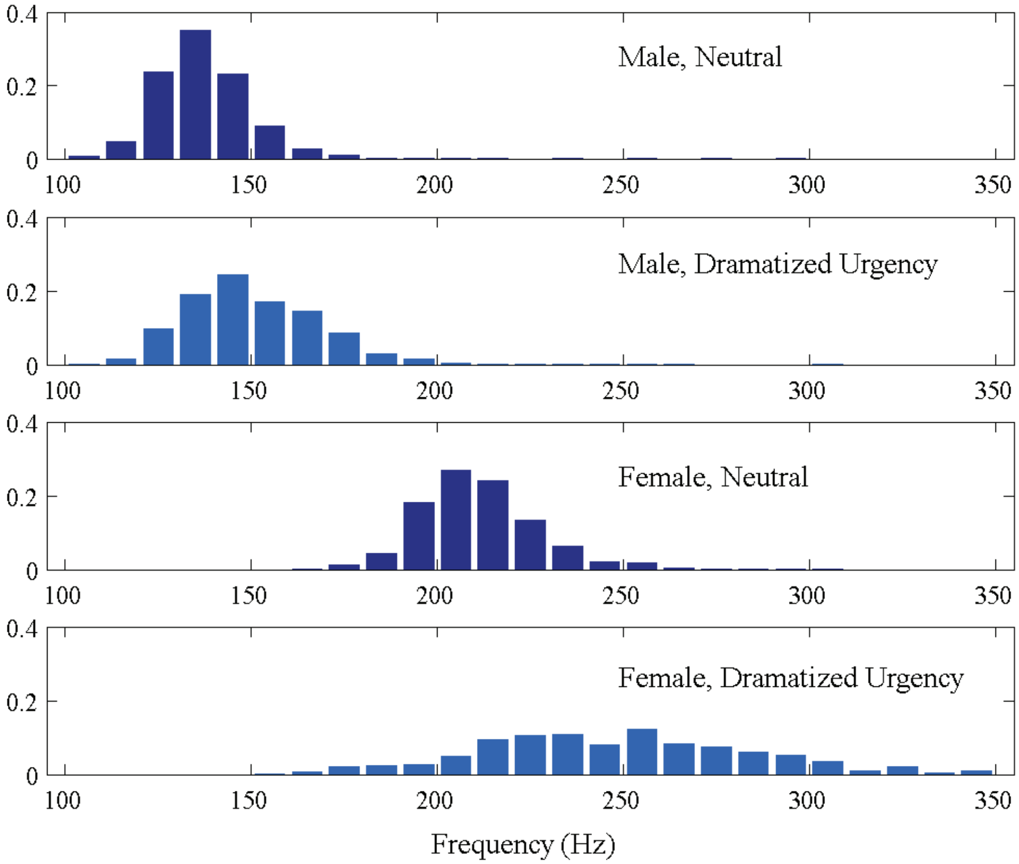


Figure 1. Pitch histograms for four cases as labeled.

The increases in mean pitch and pitch variation found in our DU recordings are qualitatively consistent with those found in cockpit voice recordings of a real stressful and urgent situation. These recordings document the voices of a pilot and copilot both when relaxed, and in the minutes before their aircraft crashes [14]. Whether or not DU is a good surrogate for true urgency will likely depend on numerous factors including individual speakers' physical and psychological characteristics and the details of the urgent situation.

All recordings for all experiments were resampled to a rate of 8,000 samples per second using the “PCM filter” option (160 to 3640-Hz bandpass filtering) provided in [16]. The level of each recording was then normalized to –26 dB, relative to clipping using tools from [16]. Next, the recordings were passed through software to implement various speech processing conditions.

4 SPEECH PROCESSING CONDITIONS

The goal of the experiments was to find the relationships between speech intelligibility, speaker identification, and the detection of dramatized urgency. The usefulness and robustness of these relationships is greatest when they span the widest possible range of these parameters. Thus, six speech processing conditions were chosen to provide the widest possible range of experimental results. Table 1 summarizes the six conditions.

Table 1. Six Conditions Used in the Speech Intelligibility, Speaker Identification, and Detection of Dramatized Urgency Experiments

Condition (C)	Description
C1	Null (no further processing)
C2	Low rate speech coding
C3	Very low rate speech coding
C4	MNRU, Q = 6 dB SNR
C5	Low rate speech coding with bit errors
C6	C5+Severe Packetization Impairments+C5

C1 involves no additional processing and thus provides a best-case reference point for all three tasks. In C4, Modulated Noise Reference Unit (MNRU) [19] software produces multiplicative (speech-correlated) noise resulting in an active speech SNR of 6 dB. This is a standardized reference condition that can allow one to build relationships to other experiments that also include the MNRU.

The remaining conditions use three different narrowband (4-kHz nominal) speech codecs specified in standards or proposed standards for low bit-rate digital communication in the presence of acoustic background noise. These codecs simulate frequency-dependent voicing strength by adaptively mixing periodic and aperiodic excitation signals. Since communications are increasingly likely to travel through some IP network it is important to model and examine possible effects. For C6, three simulated communication systems are concatenated. The first and last are the same as C5 (speech encoding, bit errors in the transmission channel, then speech decoding). The middle system consists of packetization of the speech samples followed by the deletion of randomly selected packets and the insertion of an equal number of empty packets at different random locations. A packet loss concealment algorithm is used to extend previous speech samples into these inserted empty packets.

The speech processing conditions are certainly relevant to public safety speech communication systems. But evaluating the conditions is not the primary goal of these experiments. Rather the conditions are tools that enable the experiments to yield relationships between speech intelligibility, speaker identification, and the detection of dramatized urgency.

After creating recordings for each condition, the active speech level of each recording was again normalized to –26 dB relative to clipping.

5 EXPERIMENT DETAILS

This section describes how each experiment was conducted. Evaluation of speech intelligibility, SID, and the detection of DU each required separate laboratory procedures. Each experiment was automated and used a different user interface. The speech intelligibility experiment and the detection of DU experiment used the same listeners—their time in the lab was split between the two experiments. The SID experiment was conducted about six months later, and used a different set of listeners.

5.1 Speech Intelligibility

Twenty-four randomly-selected listeners participated in the experiment. Sixteen were male, eight were female, estimated ages ranged from 20's to 60's with a mean estimated age of approximately 40, all were fluent in English, two reported slight hearing losses, and none were familiar with the technical details of the experiment. Listeners participated one at a time in a sound-isolated room where the average background noise level was below 20 dBA. The recordings were played over a powered monitor speaker with a single full-range four-inch driver. Listeners could adjust the listening level to their preferred level at any time.

In the experiment, listeners heard a recorded sentence and were asked to repeat it back. These responses were recorded and later evaluated for the number of correct words repeated. Listeners could not proceed until the entire sentence was played, and they were not allowed to replay any sentence. Progress through the experiment was controlled through a graphical interface on a PDA supported by a wireless LAN connection.

The experiment started with a practice session containing four trials. This session familiarized listeners with their task and with the procedures. Following this practice session, each listener then heard 24 sentences (4 per condition) and the sentences used with each condition were varied in a balanced way as the experiment progressed. The result was 96 intelligibility trials per condition (each sentence was used 4 times per condition, but only once per listener), for a total of 576 trials. Each listener heard the recordings in a different random order. After the experiment, several different statistical tests showed that no single listener was an outlier in this speech intelligibility task. Responses were recorded and later evaluated to compute the number of correct words repeated.

A second version of this experiment was later given to six additional listeners. This version used the same speech recordings, speech processing conditions, and general procedures. It differed from the original experiment only in that listeners were allowed to hear the recordings as many times as they wished. After each playing of a recording, listeners were asked to repeat the sentence as they heard it or to report that no words were understood. Responses were recorded and later evaluated for three quantities: the number of correct words repeated after the first playing, the number of correct words repeated after the final playing, and the total number of plays. On some occasions, a subject would fail to provide any response after a playing and this trial was scored as zero words correct.

This second version of the intelligibility experiment does not conform with typical approaches, but it does more closely parallel the SID and detection of DU experiments. In each of these, listeners were allowed to play recordings as many times as they wish.

5.2 Speaker Identification

The SID experiment design and procedures were refined several times using feedback from subjects who participated in early versions of the experiment. The final design included seven different parts. Three were actual experimental test sessions where data is collected, and four were supporting parts that were preliminary or tutorial in nature.

5.2.1 Listeners

Twenty-five randomly selected listeners participated in the experiment. Fifteen were male and ten were female. Their ages ranged from approximately 37 to 64 with a mean age of 49. None of the listeners were familiar with the technical details of the experiment. Listeners participated one at a time in a sound-isolated room where the average background noise level was below 20 dBA. Recordings were played over a powered monitor speaker with a single full-range four-inch driver. Listeners could adjust the listening level to their preferred level at any time throughout the experiment. The experiment, including all training and testing, took listeners from 45 to 90 minutes to complete, and the average completion time was just under one hour.

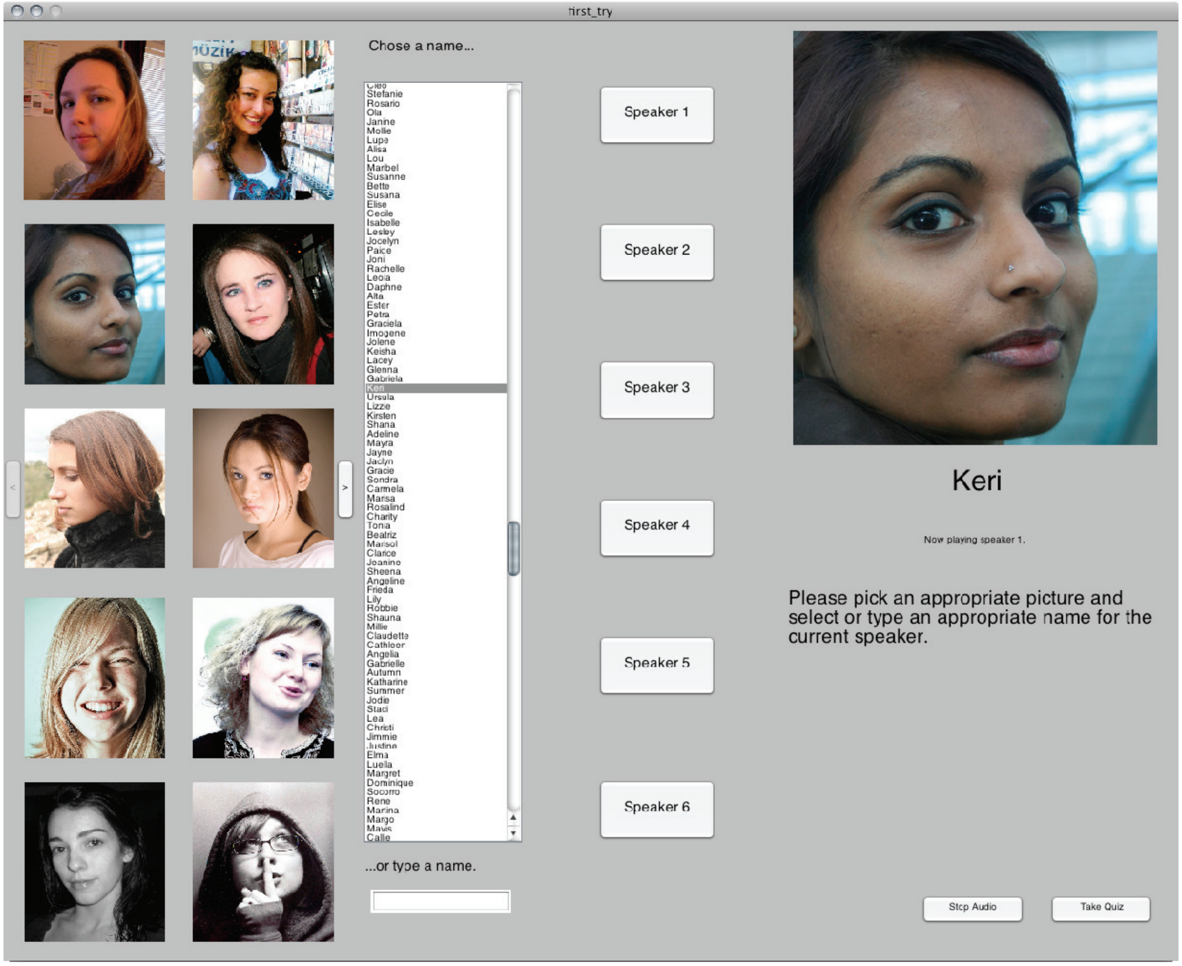
The randomly selected listener pool included two listeners with hearing aids and one listener who reported deafness in one ear. After careful consideration described in Section 6.2.1, we elected to include the data from these three listeners in the overall experiment results.

The experiment administrator received many hours of exposure to both undistorted and distorted recordings from the six speakers. After this incidental training, the experiment administrator also served as a listener. As described in Section 6.2.1, his results are not included in the overall experiment results.

5.2.2 Training and Quiz Session

The experiment starts with a session where the listener assigns a face and a name to each of the six different speakers. This session is provided to allow the experiment to better simulate the actual conditions under which listeners most often identify speakers that they cannot see. That is, listeners typically can reference a name, face, or both in memory when identifying speakers who cannot be seen.

To accomplish this goal, the SID experiment was controlled interactively through a computer program. Figure 2 shows the window presented during listener training. The middle of the window on the computer screen has one button for each speaker (e.g., “Speaker 1”, “Speaker 2”). The listener presses a button associated with a speaker. The program plays several sentences spoken by this speaker, and allows the listener to select a name and portrait[2] (or identity) that seems appropriate for the speaker from the options listed on the left side of the screen. From now on, in addition to playing several sentences spoken by this speaker, the selected identity is displayed on the right side of the screen whenever the listener presses this speaker's button.



Keri

Now playing speaker 1.

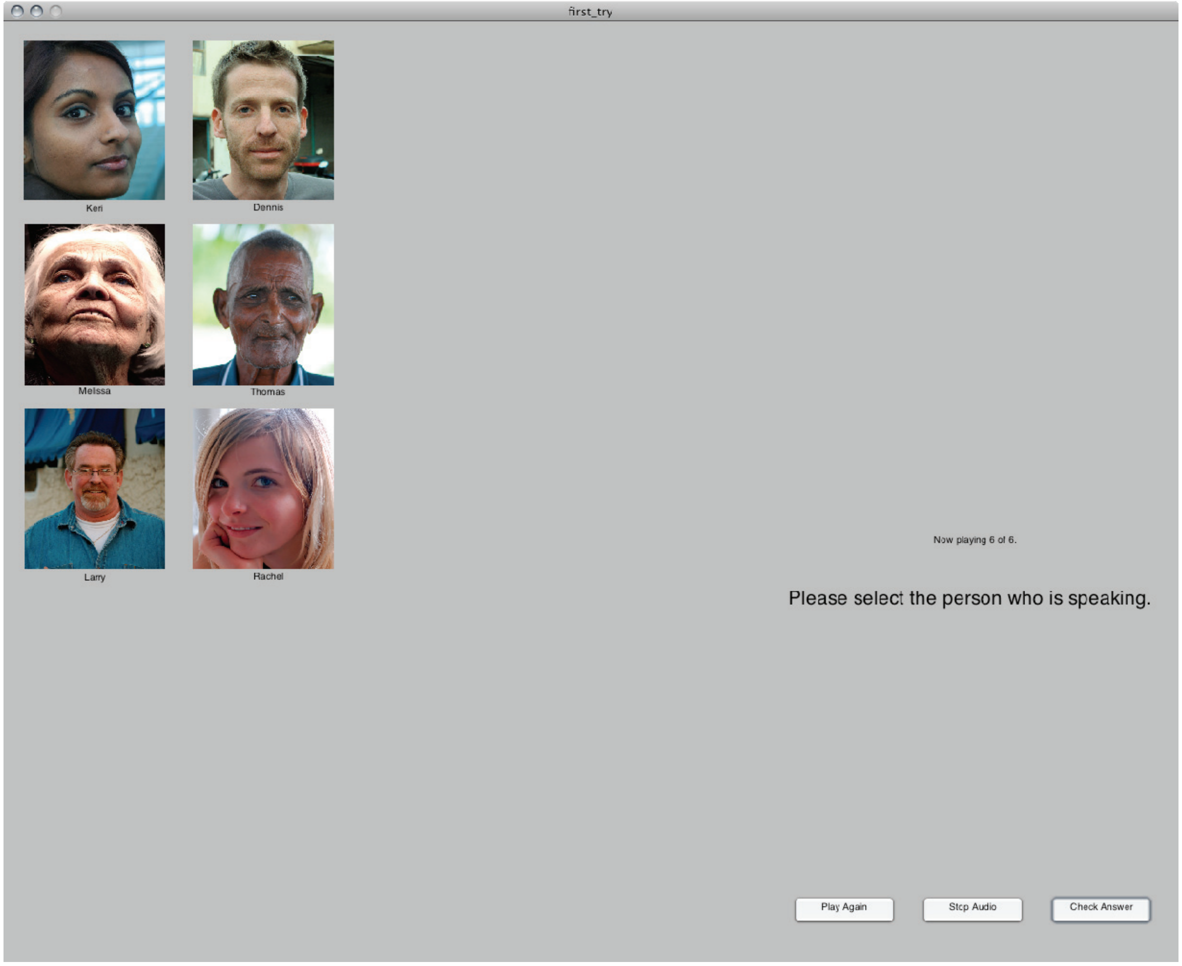
Please pick an appropriate picture and select or type an appropriate name for the current speaker.

Figure 2. Speaker identification training interface for listeners.

After hearing all six speakers and assigning a name and portrait to each, listeners take a short quiz. This quiz helps the listener check their knowledge of the speakers and become confident in their ability to recognize each speaker. The initial quiz state is shown in Figure 3. During the quiz, the listener hears a sentence. The listener identifies the speaker by clicking on the identity that he believes belongs to the currently playing speaker. The program immediately displays the selected identity prominently on the right side of the screen, as shown in Figure 4. If satisfied with the selection, the listener can confirm his choice; if not, he can make another selection. Upon confirmation, feedback indicating whether or not his selection is correct is displayed, as

[2] The portraits shown in Figures 2-5 were used under Creative Commons license; attributions listed in the appendix.

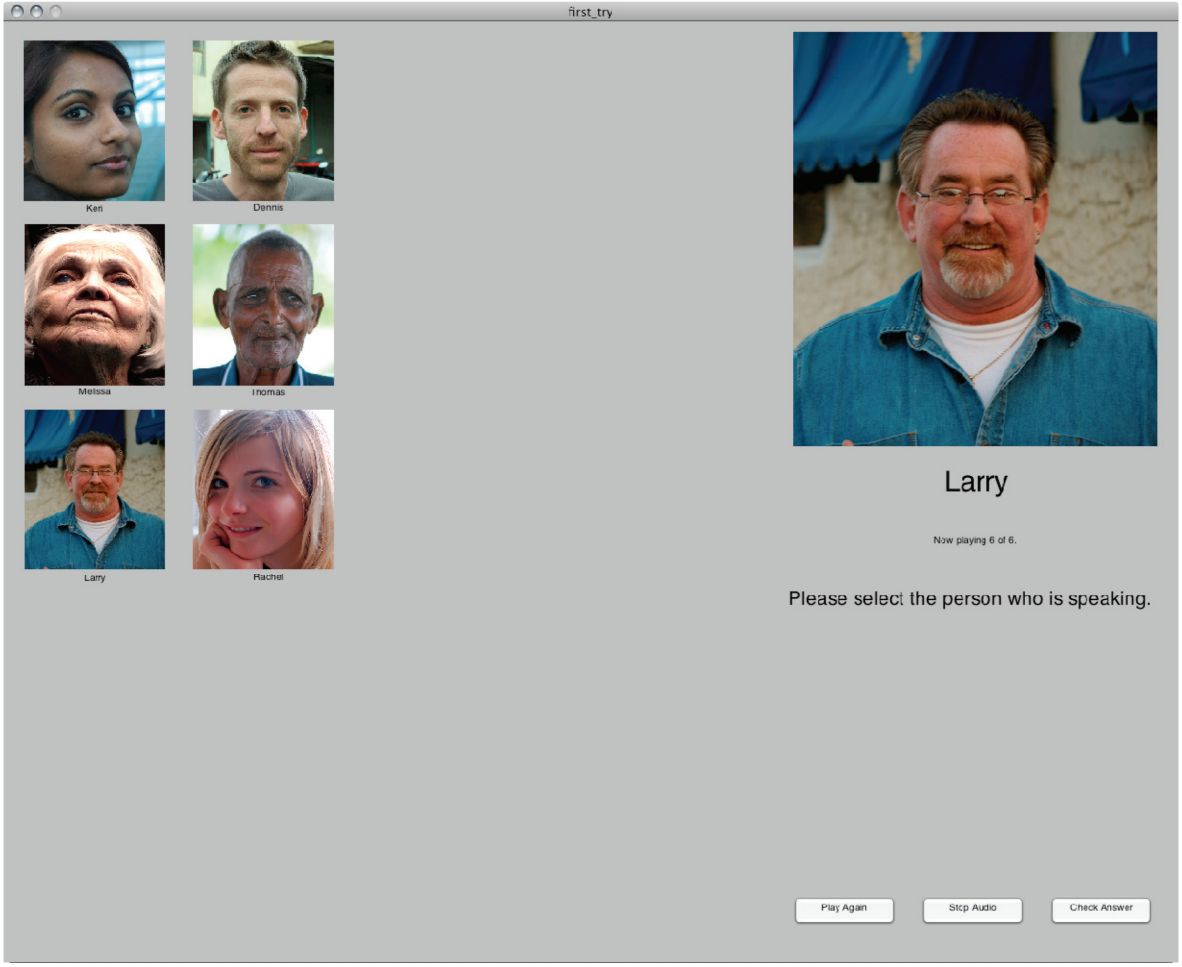
shown in Figure 5. The process repeats for all six speakers. Then the listener chooses to either continue training or proceed to the test sessions.



Now playing 6 of 6.

Please select the person who is speaking.

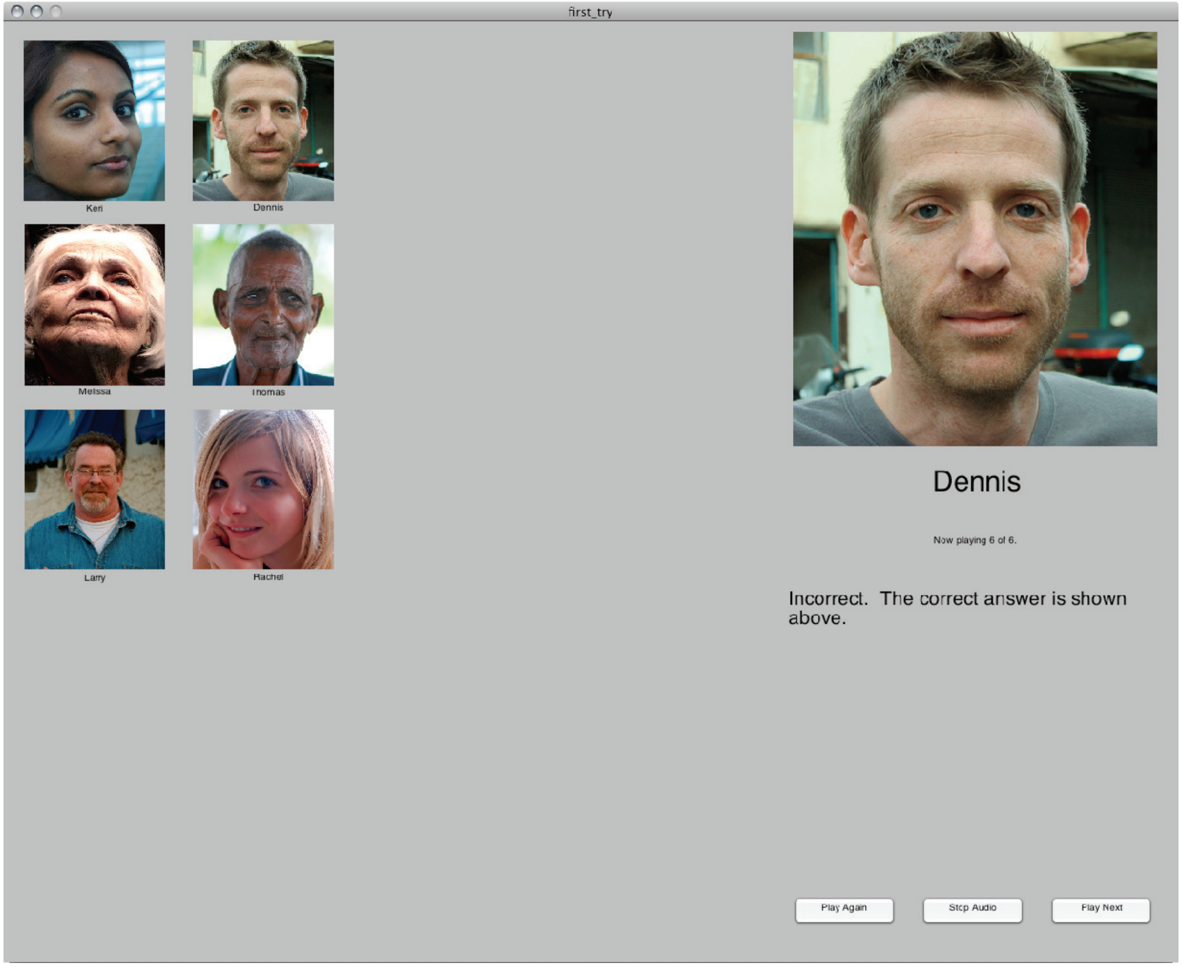
Figure 3. Speaker identification quiz interface for listeners—initial state.



Now playing 6 of 6.

Please select the person who is speaking.

Figure 4. Speaker identification quiz interface for listeners—after listener selection.



Now playing 6 of 6.

Incorrect. The correct answer is shown above.

Figure 5. Speaker identification quiz interface for listeners—correct answer display.

5.2.3 Experimental Test Session 1—Sentences

The first experimental test session uses two sentences from each speaker. There are six speakers and six conditions for a total of 2 × 6 × 6 = 72 trials in the session. One sentence is the same for all six speakers: "Don't ask me to carry an oily rag like that." The second sentence differs for every speaker. The recordings used in this session range from approximately 1.7 to 2.6 seconds in length (8 to 13 syllables in length) with a mean value of about 2.2 seconds (about 11 syllables).

This experimental test session was presented to the listener in a fashion very similar to the aforementioned quiz session. One of the 72 available recordings was played back from the beginning of a randomized list. The randomized list was unique for each listener to prevent any potential order effects. The listener was asked to identify the speaker of the recording, and select the correct identity out of the six shown on the left side of the window, on a screen similar to Figure 3. Once clicked, the selected identity displayed prominently, similar to Figure 4, and the listeners were allowed to move on to the next recording. However, the listeners were allowed to select a different identity or replay the recording as many times as necessary. Unlike the quiz session, the listeners were not notified about the accuracy of their selection—the software simply moved on to the next recording in the randomized list after identity selection was confirmed.

5.2.4 Experimental Test Sessions 2 and 3—Digits

A short reminder session is provided before experimental test sessions 2 and 3. The six chosen identities are displayed on the left side of the window. The listener is instructed to listen to each speaker at least once before moving on to the next experimental test session. By clicking any of the portraits, the listener can hear a recording of the corresponding speaker that is similar to those in the upcoming session. The listener can spend as much time in this reminder session as is desired, and the software requires that the listener listen to each speaker at least once. Once the reminder session is complete, experimental test sessions 2 and 3 are administered exactly as session 1 was.

The second experimental test session uses recordings that contain four spoken digits (e.g., "three six nine eight"). Each of the six speakers was associated with four recordings (of four digits each), so the session contained a total of 4 × 6 × 6 = 144 trials. The recordings in this session range from approximately 1.3 to 1.9 seconds in length (4 to 5 syllables in length) with a mean value of about 1.6 seconds (about 4.4 syllables). There were 15 unique "4-digit" sequences—we did not control the recording of this database, so within the speakers only 15 unique combinations were available. With one exception, all speakers recorded a completely different sub-set of these 15 sequences. Pairs of these 4-digit sequences often have two or three digits in common. Because of these similarities and subtle differences, content-based SID would be extremely difficult (e.g., noticing which 4-digit sequences are associated with Speaker 1).

The third session uses recordings that contain pairs of spoken digits. All six speakers provided the exact same four recordings: "five two", "six zero", "six three", and "eight zero". Thus, in this session, content is identical across speakers, and content-based SID is not possible. Here again, the session includes 144 trials. The recordings used in this section range from approximately 0.6 to 0.8 seconds in length (2 to 3 syllables in length) with a mean value of about 0.7 seconds (about 2.5 syllables).

The combined number of trials for all three experimental test sessions is 72 + 144 + 144 = 360.

5.3 Detection of Dramatized Urgency

This experiment used the same twenty-four randomly-selected listeners that participated in the speech intelligibility experiment (see Section 5.1). Recordings were played over a powered monitor speaker with a single full-range four-inch driver. Listeners could adjust the listening level to their preferred level at any time. Experiment progress and data collection were controlled through a graphical interface on a PDA supported by a wireless LAN connection. Listeners first participated in a practice session to familiarize them with the task and the procedure.

Listeners heard a recording and responded to the prompt "Please select the talker's stress or urgency level." Response options in each of these binary forced-choice trials were "Low" (the correct answer for neutral recordings) and “High” (the correct answer for TIS and DU recordings). Listeners could respond at any time once a recording had started to play, and could restart the playback at any time. In this manner, each listener could proceed at an individualized pace through the experiment. Each listener heard 192 trials for a total of 4,608 (192 trials × 24 listeners) DU detection data points.

6 RESULTS

Throughout this section and the next, we report results in terms of either fraction correctly identified or normalized task performance (NTP). We introduce the NTP scale because it enables a more direct comparison of the results from the three experiments. On the NTP scale, a value of 1 indicates perfect information from the listeners in the experiment, and a value of 0 indicates no information from the listeners. This is true for the intelligibility, SID, and detection of DU experiments.

6.1 Speech Intelligibility

In the intelligibility experiment, NTP is simply the fraction of words correctly repeated by a listener. If 100 percent of the words were repeated correctly, an NTP value of 1 would be the result. If none of the words were repeated correctly, then an NTP value of 0 would be the result.

Figure 6 shows the mean NTP values for each of the six speech processing conditions, after averaging over all listeners and all messages for each condition. The data points shown by the solid line describe the main intelligibility experiment (24 listeners, replaying recordings not allowed). The data points shown by the dashed and dotted lines correspond to results from the second version of the intelligibility experiment (six listeners, replaying recordings allowed). The dashed line shows results that were calculated after the first play and the dotted line shows results that were calculated after the final playing. The 95-percent confidence intervals (CI) for these three cases overlap and are omitted from this figure for clarity.

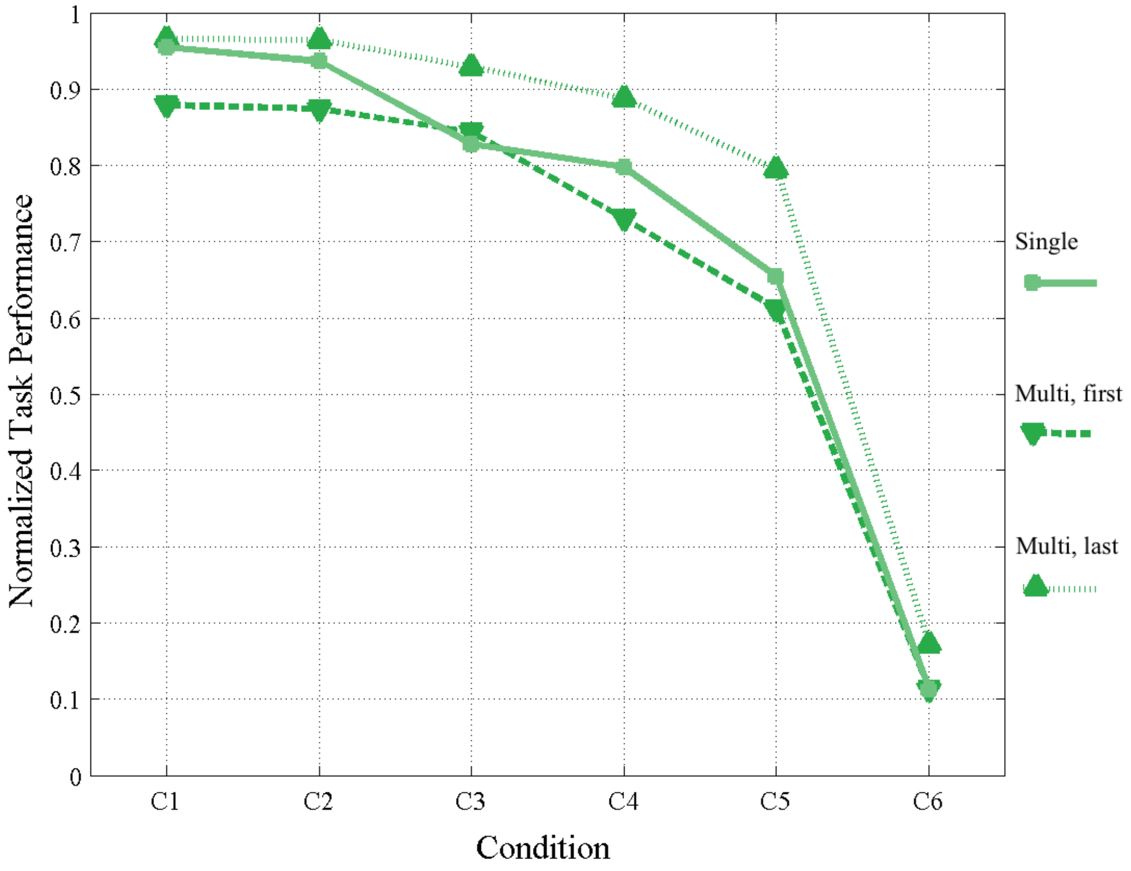


Figure 6. Intelligibility experiment: NTP values for three cases—solid line for single play experiment, dashed line for the first play, and dotted line for the last play in the multiplay experiment.

Figure 6 shows that allowing replays does improve NTP by a small amount (the dotted line is above the solid line). The larger increases in NTP are associated with cases of medium speech intelligibility (C4 and C5) where additional plays apparently can help at least some listeners with the task. The smaller increases in NTP are associated with cases of very high (C1, C2, and C3) and very low intelligibility (C6). It seems that in these limiting cases, additional plays provide limited advantage. The average number of plays generally increases with condition number from just over one (1.3 in C1) to two (2.0 in C6).

While allowing replays does improve NTP by a small amount this does not change the general trends as we move from C1 to C6. That is, all three lines in Figure 6 show a similar general trend: a gentle decrease in NTP followed by an abrupt drop. Based on this similarity of shape we argue that the final conclusions about relative robustness in Section 7 are not highly sensitive to the choice between these two intelligibility testing approaches (single play of each recording versus unlimited playing of each recording). The main intelligibility experiment (with no replays allowed) has much more data than the secondary intelligibility experiment (24 listeners vs. 6 listeners) and thus we focus on that experiment from this point forward in this report.

Figure 7 shows the mean NTP values and 95 percent CI for each of the six speech processing conditions, after averaging over all listeners and all messages for each condition. Note that as we move from C1 to C6, NTP drops monotonically from 0.95 to 0.11. This is an NTP drop of 0.84.

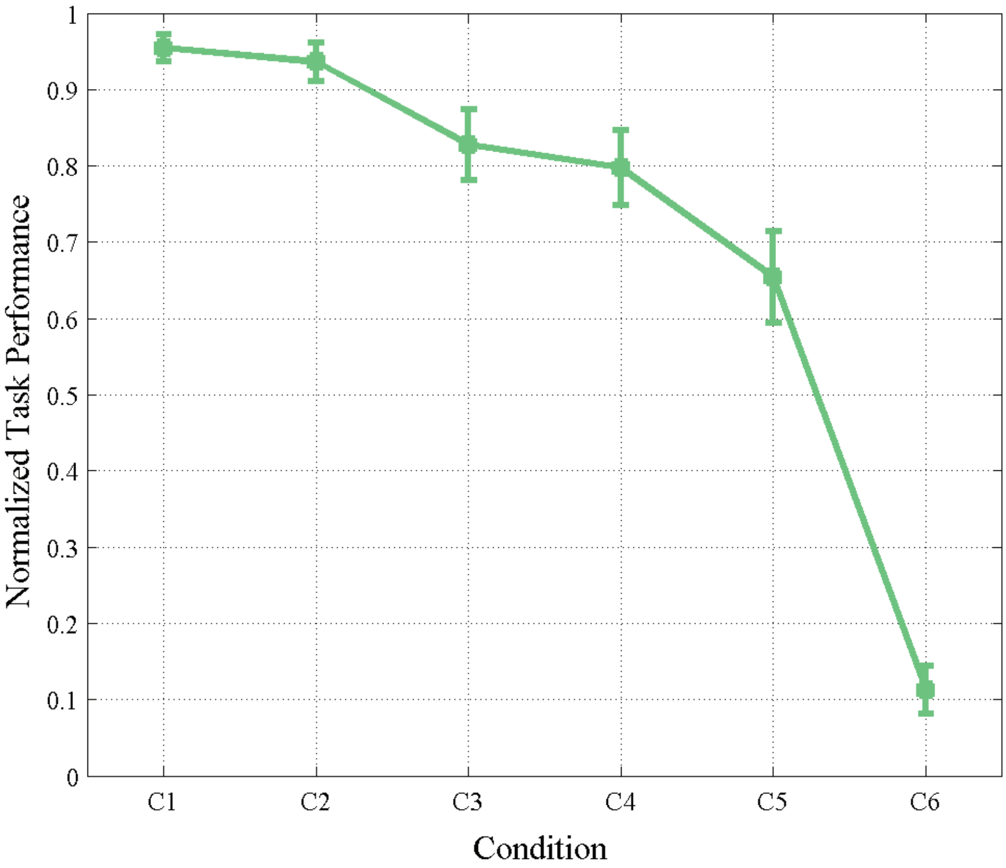


Figure 7. Intelligibility experiment: NTP mean and 95% CI for word intelligibility in sentence context.

6.2 Speaker Identification

In the SID experiment, 360 trials were administered to 25 listeners in the main pool. This gives 9,000 data points. Each data point is one SID, which can be either correct or incorrect. Using this view, the data is binary in nature and can be modeled using the binomial distribution. In the binomial model, the maximum likelihood estimate for the probability of correct identification is simply as shown in (1).

$$\hat{p} = \frac{number\ of\ correct\ \ identifications}{total\ number\ of\ identifications} \tag{1}$$

The 95% CI for the estimate $\hat{p}$ is calculated as given in [18]. We report $\hat{p}$ as the fraction correctly identified in Section 6.2.1 and Section 6.2.2, and we report the 95% CI for the estimate $\hat{p}$ as well.

Sections 6.2.1 through 6.2.3 discuss the results of the SID experiment in three different ways—broken up by listeners, speakers, and then more directly by speech processing condition. When the data is broken up by listener, it is clear that each listener scored better than would be likely if he was randomly guessing. While there was a general decreasing trend in NTP, there were no statistically significant differences among sessions. Therefore, results broken up by session (or clip length) are not discussed. Some speakers are easier to recognize than others, and those not easily recognized are often confused with other speakers.

6.2.1 Listeners

Figure 8 shows the fraction of correct identifications and the associated 95% CI for the 25 listeners. Listeners are numbered and sorted by increasing fraction of correct identifications. The mean fraction of correct identifications over all listeners is 0.662, and 20 of the 25 listeners have overall correct identification fractions between 0.59 and 0.81. The people who utilized hearing aids have listener numbers 14 and 16. Listener 16 was also a non-native speaker. The listener who reported deafness in one ear is listener number 20. Figure 8 shows that none of these three listeners is an outlier. Thus, all three are retained in our data pool.

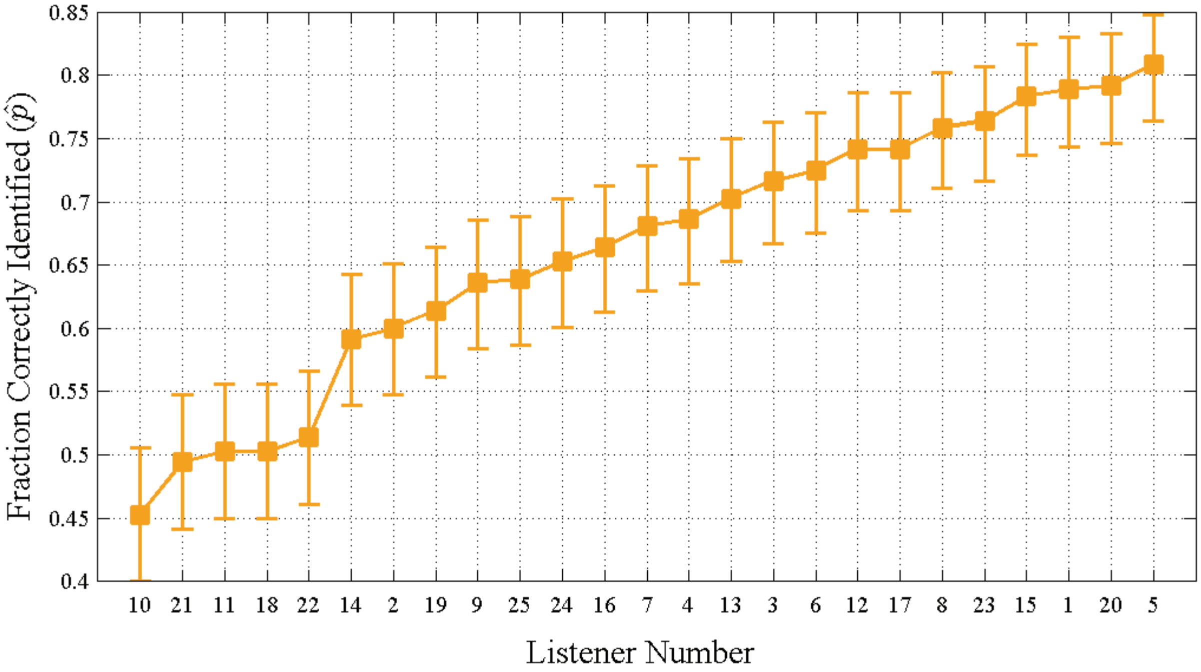


Figure 8. SID experiment: Mean fraction correctly identified by listener and 95% CI for each listener.

Out of the three listeners whose first language was not English, only one seemed to be at a disadvantage (listener 10). The other two non-native English speakers (listeners 13 and 16) placed close to the fraction-correct mean among all listeners. We elected to retain all three of these listeners in our data pool.

Not shown in Figure 8 is the experiment administrator, who received a great deal more training (more than 20 hours on speech distorted under all conditions). The experiment administrator was significantly more accurate with a 0.98 fraction of correct identifications. This is an indication that additional training can have a positive effect on SID performance. It's possible that the experiment administrator memorized to whom many of the recordings belonged, but it's also possible that over time he learned how each condition affected speech. The latter possibility may be relevant, as it is possible for listeners to become accustomed to adverse communication conditions. Once again, the experiment administrator's results were not included in the overall experiment results.

6.2.2 Speakers

Our selection of speakers had some interesting properties. The males had average pitches of 92, 105, and 111 Hz, and Male 3 had a slight Southern accent. The females had average pitches of 103, 104, and 107 Hz. Female 1 had a Midwestern accent, Female 2 had a Southern accent and Female 3 had a heavy Ecuadorian accent. The task of distinguishing among the three females is made easier (relative to the task of distinguishing among the three males) by very pronounced accents despite their small average pitch spread relative to that of the males.

Table 2 is the confusion matrix for the SID task for these six speakers averaged across all clips, conditions, and listeners. Each row is associated with one speaker, and each column is associated with the listener votes. "M" indicates male, "F" indicates female. Shaded cells indicate the fraction of correct SID, unshaded cells indicate fractions of confused SID. For example, the top left entry indicates that 67 percent of the clips from Male 1 were identified as coming from Male 1. The next entry to the right indicates that 22 percent of the clips from Male 1 were identified as coming from Male 2. Similarly, the next entry to the right indicates that 11 percent of the clips from Male 1 were identified as coming from Male 3.

Table 2. SID Experiment: Confusion Matrix

		Listener Selected Speaker					
		M1	M2	M3	F1	F2	F3
Actual Speaker	M1	0.670	0.220	0.110	0.000	0.000	0.000
	M2	0.150	0.570	0.220	0.010	0.030	0.030
	M3	0.120	0.340	0.540	0.000	0.000	0.000
	F1	0.000	0.003	0.001	0.650	0.190	0.160
	F2	0.000	0.004	0.001	0.170	0.740	0.080
	F3	0.001	0.003	0.005	0.070	0.120	0.800

The confusion matrix shows that Female 3 is easier to identify than Female 2, who in turn is easier to identify than Female 1 (correct identification fractions of 0.80, 0.74, and 0.65, respectively). Male 1 (with a correct identification fraction of 0.67) and Female 1 are close to the same difficulty, and Males 2 and 3 (fractions of 0.57 and 0.54, respectively) are both more difficult. The task of distinguishing among the males is difficult because Males 2 and 3 sound very similar (despite a slight Southern accent present in Male 3). In fact, the matrix shows that the greatest levels of confusion are between Males 2 and 3, though confusions between Male 1 and Male 2, and confusions between Female 1 and Female 2, are not far behind.

As Schmidt-Nielsen notes, listeners perform the SID task more efficiently with familiar, or distinctive speakers [6]. Our results are consistent with prior research. The two male speakers who had the smallest fraction of correct identifications were also often expressed as perceptually similar by listeners during the experiment. While the average pitch difference between the two easily confused male speakers is greater than the pitch spread among all female speakers, the female speakers were arguably more distinctive due to their regional accents. The speech processing conditions used in these experiments preserve pitch in the general sense, though the finer nuances of pitch trajectories will sometimes be lost.

Listeners received 15 minutes of training with these 6 unfamiliar speaker voices on average. Listeners were allowed to continue the training process as long as they wished. Many of the SID trials involved recordings in which the voice was greatly distorted. Thus, this amount of confusion is not unexpected. It is interesting to note that only Male 2 is ever perceived to be a female; all three females are confused for males, but only rarely.

The difficulty of the SID task is broken down by speaker and by condition in Figure 9. Males 1, 2, and 3 are shown with red circles, orange stars, and dark gray downward-pointing triangles and lines, respectively. Females 1, 2, and 3 are shown with blue upward-pointing triangles, green squares, and purple diamonds and lines, respectively. With few exceptions, easier-to-identify speakers tend to be easier for all six conditions, and harder-to-identify speakers tend to be harder for all six conditions. The major exception is female 1 who is one of the easiest-to-identify speakers when heard over C1, C2, and C4, but is one of the hardest-to-identify speakers when heard over C5 and C6.

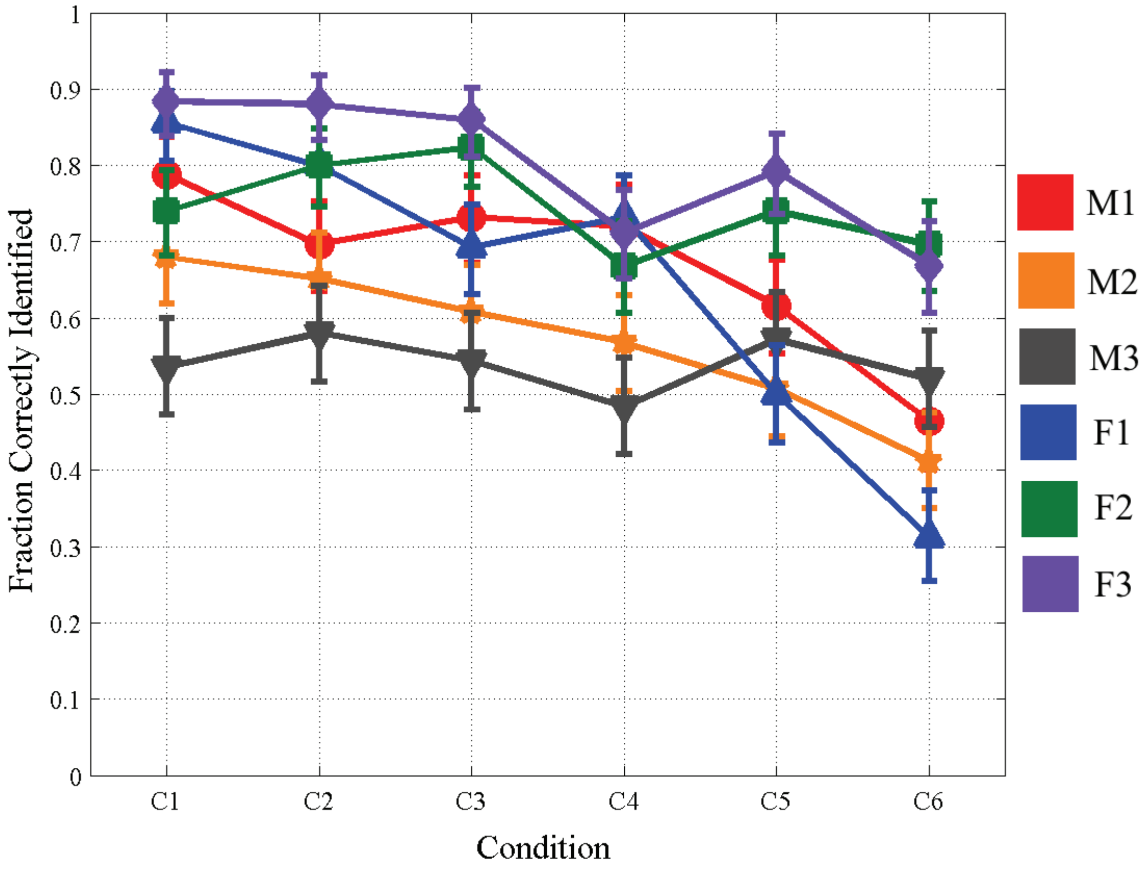


Figure 9. SID experiment: Fraction correctly identified and 95% CI by speaker and condition.

6.2.3 Speech Processing Conditions

A main goal of this work is to quantify how the listener SID performance is influenced by the six speech processing conditions. Each of the six conditions described in Table 1 was used for a total of 1,500 SID trials. For each condition, these 1,500 trials used the same 60 recorded speech files, and the same 25 listeners as well. This balance allows us to compare SID results for the six conditions directly, as Figure 10 shows.

For SID, the transformation from estimated probability of correct identification $\hat{p}$ to NTP is shown in (2).

$$\mathrm{NTP} = \frac{6}{5} \times \left(\hat{p} - \frac{1}{6} \right), \qquad \frac{1}{6} \le \hat{p} \le 1 \tag{2}$$

Because six responses are possible in this experiment, a listener making no effort and giving strictly random responses could have an average fraction of correct identifications of $\frac{1}{6}$. Thus, $\frac{1}{6}$ corresponds to no information from a listener, and (2) maps $\frac{1}{6}$ to an NTP value of 0. On the other hand, perfect SID corresponds to an NTP value of 1.

The transformation given in equation 2 is not defined for values of $\hat{p}$ that are less than $\frac{1}{6}$ (worse than guessing). While such values are theoretically possible, a meaningful interpretation is not obvious and would likely depend on additional factors. In this experiment no listeners exhibited performance worse than guessing. We expect that such results would be an extremely rare aberration and thus we do not consider the restriction in equation 2 to be a significant limitation.

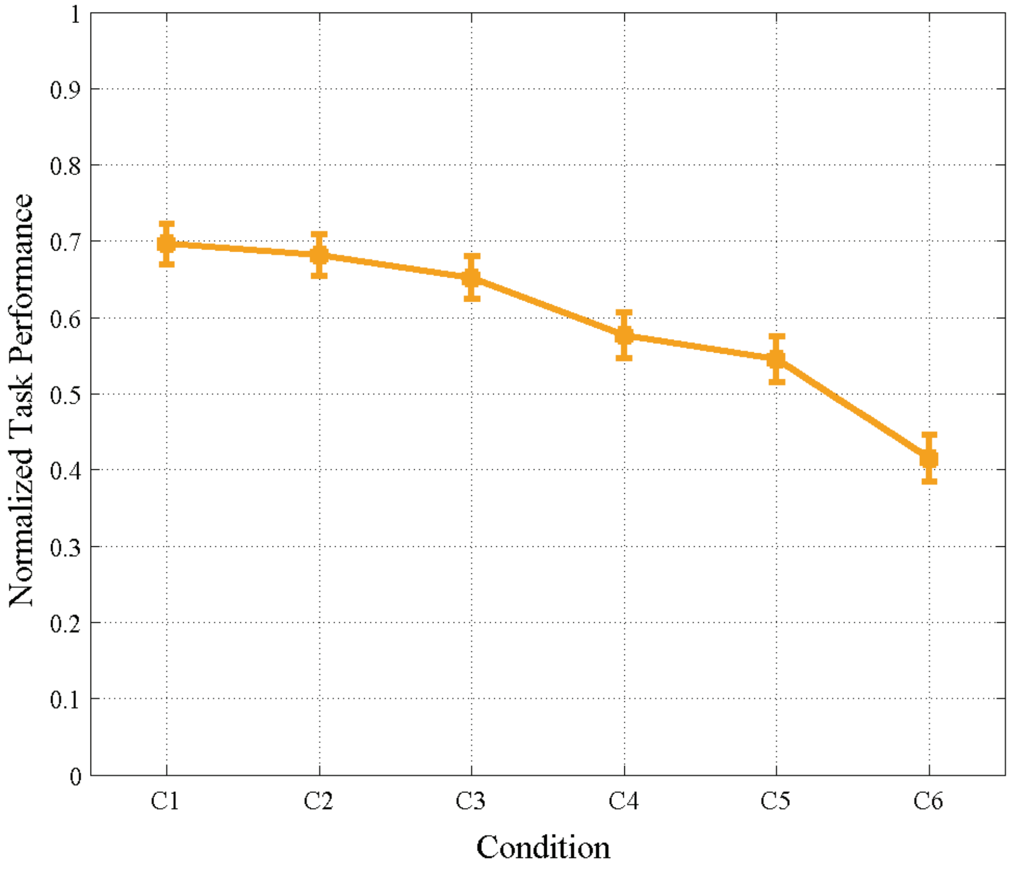


Figure 10. SID experiment: NTP mean and 95% CI by condition.

Note that as we move from C1 to C6, the SID NTP drops steadily from 0.69 to 0.41. This is an NTP drop of 0.28.

6.3 Detection of Dramatized Urgency

For each trial in a detection experiment, three outcomes are possible: correct detection, false alarm (low urgency reported as high urgency), and miss (high urgency reported as low urgency). Given the binary nature of the data (correct or not correct), the maximum likelihood estimate for the probability of correct detection and the 95% CI for that estimate are calculated as given in (1) and [18]. As with SID, we report detection of DU results in terms of the NTP scale. In this case, that scale is defined by (3).

$$\text{NTP} = 2 \times \left(\hat{p} - \frac{1}{2} \right) \tag{3}$$

Because two responses are possible in this experiment, a listener making no effort and giving strictly random responses could have an average fraction of correct identifications of $\frac{1}{2}$. (3) maps this to an NTP value of 0.

Figure 11 shows the mean NTP values and 95% CI for each of the six speech processing conditions, after averaging over all listeners and all messages for each condition. Figure 11 shows that as one progresses from C1 to C6, the NTP for detection of DU in messages drops from 0.76 to 0.58 (an NTP drop of 0.18). We also found that across the conditions, the false alarm rate tends to be lower than the miss rate. The false alarm rates generally fall into the range 0.05 to 0.10, while the miss rates generally span the range 0.10 to 0.35. In other words, detection-of-DU errors are less frequent when speakers are in the neutral state, and more frequent when speakers are in the DU state.

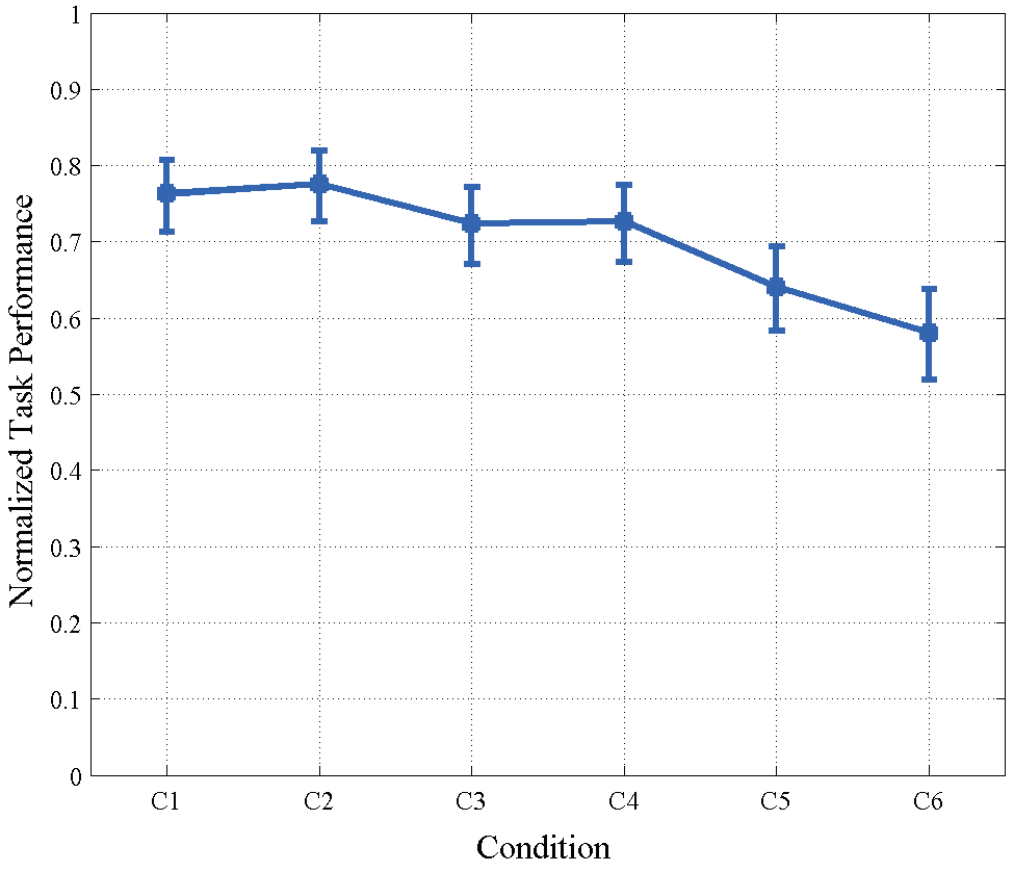


Figure 11. Dramatized Urgency Experiment: NTP mean and 95% CI for detection of dramatized urgency.

7 COMBINED RESULTS AND DISCUSSION

This section contains the results from 6.1, 6.2.3, and 6.3 combined. This allows one to compare the mean intelligibility, SID, and detection of DU results across the six speech processing conditions, consistent with the overall goal of this work. These combined results are given in Figure 12. The green line shows NTP mean and 95% CI for word intelligibility; the blue line shows mean and CI for detection of dramatized urgency; the yellow line shows mean and CI for SID. All three results generally decrease as one progresses from C1 to C6.

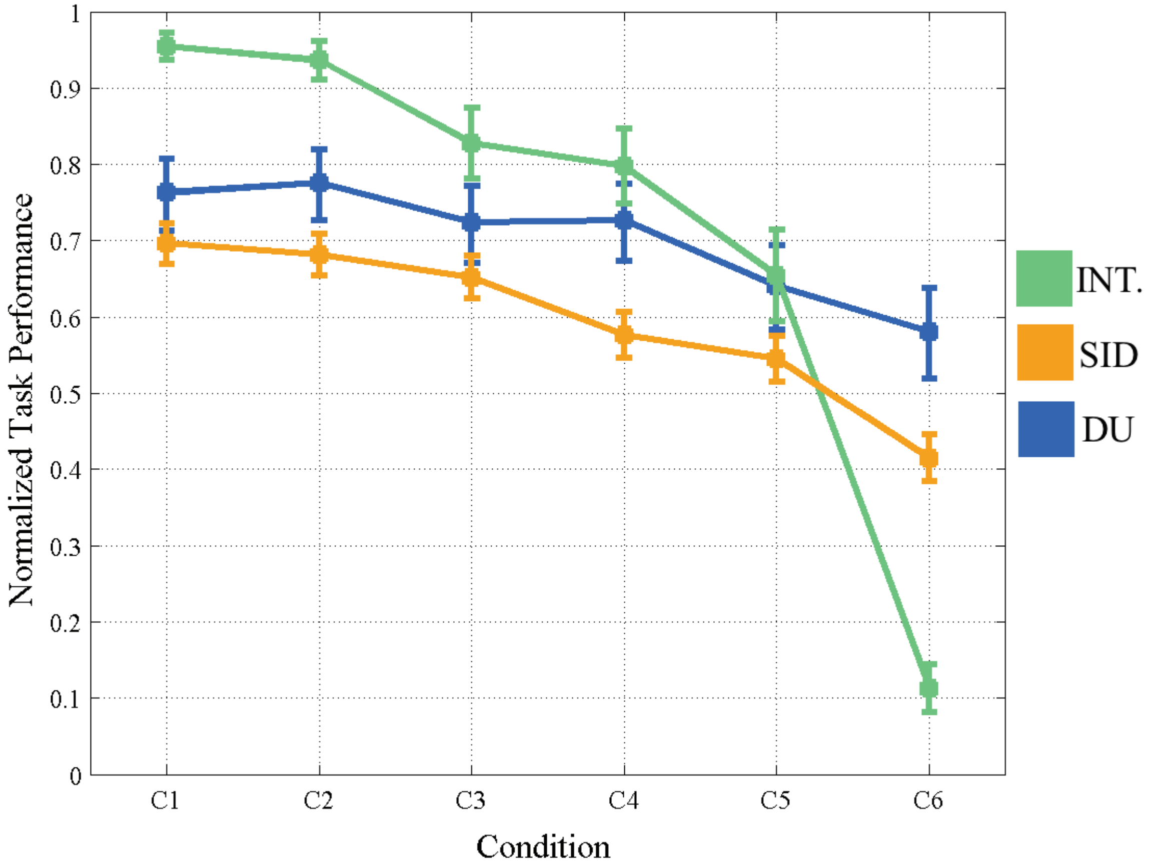


Figure 12. NTP mean and 95% CI for word intelligibility in a sentence context (green), detection of DU (blue), and SID (yellow).

In the following we compare these decreases in intelligibility, DU, and SID NTP from C1 to C6 instead of comparing the raw NTP values of those three quantities for each speech processing condition. We believe that comparing these decreases leads to conclusions that are more likely to generalize beyond this specific set of experiments. It is clear that conclusions based on the raw NTP values do not generalize beyond this specific set of experiments because the raw NTP values can be very strongly influenced by experiment design.

If the SID experiment had used 24 speakers instead of 6 speakers, that task would have been much harder and the SID results curve in Figure 12 would undoubtedly be shifted downward dramatically, likely resulting in saturation with an NTP very close to 0 for all six speech processing conditions. Conversely, if the SID experiment had used just one female and one male speaker, that task would have been much easier and the SID results curve would have undoubtedly shifted upward dramatically, likely resulting in saturation with an NTP near 1 for all six speech processing conditions. Such dramatic shifts can saturate the NTP scale and this can prevent us from finding the information we seek in this work: how are intelligibility, DU, and SID influenced by the six speech processing conditions?

Figure 12 shows that our experiment designs have been successful in that none of the three curves saturates at the top or bottom of the NTP scale. In light of the discussion above, statements about SID NTP relative to DU NTP for C3, for example, are specific to the parameters of these experiments (e.g., six talkers in SID, two levels of DU, and a host of other experiment parameters). On the other hand, an examination of the decreases in intelligibility, SID, and DU NTP as we progress from C1 to C6 leads to somewhat more general conclusions about the relative robustness of intelligibility, SID, and DU to the distortions induced by the six speech processing conditions. While these decreases in NTP may also be influenced by experiment design, we argue that they must be much less sensitive than the raw NTP values.

Figure 12 shows the intelligibility results decrease most abruptly (from 0.95 to 0.11 for a decrease of 0.84), and the detection of DU results decrease most gently (from 0.76 to 0.58 for a decrease of 0.18). The SID results show a decrease that is between the other two (from 0.69 to 0.41 for a decrease of 0.28). If we compare the NTP decrease for SID with the NTP decrease for intelligibility (0.28 compared with 0.84), we can conclude that the SID is 3.0 times (0.84/0.28) more robust to the distortions created by the speech processing conditions than intelligibility is. Similarly, comparison of the NTP decrease for detection of DU with the NTP decrease for intelligibility (0.18 compared with 0.84) indicates that the detection of DU is 4.7 times (0.84/0.18) more robust to the distortions created by the speech processing conditions than intelligibility is.

These are the final relationships to be extracted from these experiments. They suggest that if a speech communication system is well represented by the speech processing conditions used in these experiments and its distortions are such that word intelligibility NTP has not decreased significantly (a very natural requirement for any useable system) then those distortions are unlikely to have caused SID and DU NTP to have decreased significantly.

Laboratory experiments like those described in this report are important because they provide a level of control over speaking, listening, and speech processing conditions that allows one to extract meaningful results. This would not be possible in a typical field environment. While laboratory experiments are essential to research progress, it is also true that the laboratory is often less realistic than the field environment.

One factor to consider in this regard is the consequence of various types and levels of background sounds at speaker and listener locations. It is clear that these background sounds can have negative effects, but they might also aid in SID (e.g., when it is known that Officer Roberts is at the coffee shop and Officer Smith is at a subway station, the corresponding background sounds could help with SID). They might also enable detection of urgency in speakers' voices. Examples of background sounds that may indicate a situation in which an officer could be exhibiting urgency in his voice may be screeching tires or even gunfire.

The relationship between dramatized urgency and the actual emotional signatures found in the voices of public safety officials is also of great interest. Dispatchers and officials who deal with urgent, catastrophic or tragic events on a routine basis may show less emotional variation in their voices than the general public would. However, it is possible that even officials with the calmest demeanor may find themselves in situations where they would exhibit urgency in their speech. Our results show that if this were to occur, the listener would likely be able to detect the sound of urgency. If intelligibility needed to be enhanced, officials could speak more slowly and clearly, but this could have a negative effect on SID, in that the speaker's normal cadence and speaking rhythm would be altered.

An additional issue centers on SID of familiar and unfamiliar speakers. Certainly years of professional association can cause voices and speaking styles to be very familiar, even under adverse conditions. This could lead to SID rates higher than those found in this laboratory study that uses unfamiliar speakers and a relatively short training or "acquaintance" process. On the other hand, many or most public safety officials communicate with far more than six other officials on a regular basis. This could make the SID task more difficult. How these two competing effects might balance out could only be determined through additional research efforts.

We designed the DU and SID experiments to have minimal complexity and yet provide useful basic DU and SID information. It is clear that one could design a large number of more intricate, more complex, or more specific experiments to explore the issues discussed in this section and a host of other associated issues. As experiments become more complex and specific, it becomes increasingly important to pay close attention to the balance between control and realism so that experiment results are of maximum utility and applicability.

8 ACKNOWLEDGEMENTS

This work was funded by the Department of Homeland Security Office of Interoperability and Compatibility, through the National Institute of Standards and Technology Office of Law Enforcement Standards. The work was conducted at the Institute for Telecommunication Sciences under the supervision of J.R. Bratcher.

The authors of this report wish to express deep gratitude to Ryan Buffington and Alicia DeMino for implementing and executing the intelligibility and DU experiments, including database identification and acquisition, recording speech samples, software development, listener recruitment, and experiment administration.

We also wish to thank the anonymous listeners from the US Department of Commerce Boulder Labs who participated in these experiments.

9 REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, J. Taylor, and W. Fellenz, “Emotion recognition in human-computer interaction,” *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[2] S. Bou-Ghazale and J. Hansen, “A comparative study of traditional and newly proposed features for recognition of speech under stress,” *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 4, pp. 429–442, Jul. 2000.

[3] A.J. Compton, “Effects of filtering and vocal duration upon the identification of speakers, aurally,” *The Journal of the Acoustical Society of America*, vol. 35, no. 11, pp. 1748–1752, 1963.

[4] P.D. Bricker and S. Pruzansky, “Effects of stimulus content and duration on talker identification,” *The Journal of the Acoustical Society of America*, vol. 40, no. 6, pp. 1441–1449, 1966.

[5] Z. Uzdy, “Human speaker recognition performance of LPC voice processors,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 3, pp. 752–753, Jun. 1985.

[6] A. Schmidt-Nielsen and K.R. Stern, “Identification of known voices as a function of familiarity and narrow-band coding,” *The Journal of the Acoustical Society of America*, vol. 77, no. 2, pp. 658–663, 1985.

[7] A. Schmidt-Nielsen and K.R. Stern, “Recognition of previously unfamiliar speakers as a function of narrow-band processing and speaker selection,” *The Journal of the Acoustical Society of America*, vol. 79, no. 4, pp. 1174–1177, 1986.

[8] A. Schmidt-Nielsen, “A test of speaker recognition using human listeners,” in *Proc. 1995 IEEE Workshop on Speech Coding for Telecommunications*, Annapolis, Maryland, Sep. 1995, pp. 15–16.

[9] A. Schmidt-Nielsen and D.P. Brock, “Speaker recognizability testing for voice coders,” in *Proc. 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, Georgia, May 1996, vol. 2, pp. 1149–1152.

[10] A. Schmidt-Nielsen and T.H. Crystal, “Human vs. machine speaker identification with telephone speech,” in *Proc. 5th International Conference on Spoken Language Processing*, Sydney, Australia, Nov. 1998, vol. 2, pp. 221–224.

[11] T. F. Quatieri, *Discrete-Time Speech Signal Processing, Principles and Practice*, Chapter 14, Upper Saddle River, NJ: Prentice Hall, Inc., 2002.

[12] Tactical Speaker Identification Database, Available at http://www.ldc.upenn.edu.

[13] H. Steeneken and J. Hansen, “Speech under stress conditions: Overview of the effect on speech production and on system performance,” in *Proc. 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Mar. 1999, vol. 4, pp. 2079–2082.

[14] R. Ruiz, E. Absil, B. Harmegnies, C. Legros, and D. Poch, “Time-and spectrum-related variabilities in stressed speech under laboratory and real conditions,” *Speech Communication*, vol. 20, no. 1-2, pp. 111–129, Nov. 1996.

[15] Speech under Simulated and Actual Stress Database, Available at http://www.ldc.upenn.edu.

[16] ITU-T Recommendation G.191, Software tools for speech and audio coding standardization, Geneva, 2005.

[17] ITU-T Recommendation P.810, Modulated noise reference unit (MNRU), Geneva, 1996.

[18] N. Johnson, S. Kotz, and A. Kemp, *Univariate Discrete Distributions*, New York: Wiley, second edition, 1992, p. 129.

APPENDIX: PHOTO ATTRIBUTIONS

All photos used in the SID portion of this experiment were under a Creative Commons license at the time of download. Attributions are listed below. The artist's name or alias is listed first and is followed by their flickr <http://flickr.com> username in parentheses. If no alias is listed, the username of the account where the picture was downloaded at the time of download is listed.

Johannes Henseler (frischmilch flickr)

Vladislav Sabanov (willgame on flickr)

Katie Tegtmeyer (Katie Tegtmeyer on flickr)

Amy March (Amy March on flickr)

"Patrick" (LordKhan on flickr)

"Sarah G" (SpooSpa on flickr)

Igor Maminta (igorms on flickr)

(mwyllie on flickr)

"Polo" (pulpolux !!! on flickr)

Alexis Stapovic (Alexis Stapovic on flickr)

Sandra Nahdi (Concentrated Passion on flickr)

(Cristina _sanza on flickr)

"Aure" (Versatile Aure on flickr)

Caleb Sconosciuto ('SeraphimC on flickr)

Jesper Sachmann (Sachmanns.dk on flickr)

Mareen Fischinger (Mareen Fischinger on flickr)

"W2 Beard & Shorty" (W2 a-w-f-i-l on flickr)

Roland Lakis (rolands.lakis on flickr)

"Corey" (Coreu on flickr)

"Jonas K." (jonas_k on flickr)

"SSynth" (SSynth on flickr)

"Carla S." (le jardin public - CS Photo on flickr)

Simón Pais-Thomas (Simón Pais-Thomas on flickr)

(* michel clair on flickr)

(bs70 on flickr)

"CK, Carl, Carlo, Carlito" (09traveler on flickr)

allfr3d (allfr3d on flickr)

Fernando Gregory (ƒreg on flickr)

(harvey20887 on flickr)

Michael Verhoef (nettsu on flickr)

"Thomas" (streetpreacher83 on flickr)

(Pictr 30D (B2B) on flickr)

(kevin41890 on flickr)

(newpn2000 on flickr)

(ooodit on flickr)

(Idleeuw on flickr)

Chandrachoodan Gopalakrishnan (Ravages on flickr)

yatenkaiouh (yatenkaiouh on flickr)

Frédéric DUPONT (darkpatator on flickr)

(corey (a.k.a. ten0fnine) on flickr)

Felix Dylan (broma on flickr)

Jiraroj Sheravanichkul (bookazine on flickr)

shafiu Jameel (shaapay on flickr)

David Lytle (davitydave on flickr)

Tairre Christopherson (milomingo on flickr)

(laurabaabaa on flickr)

"Emiliano" (loungerie on flickr)

Adrián Flores (-.:Ferran : WWT's Ambassador to the Aztecs:.- on flickr)

"Ra'anan Niss." (Ronan_tlv on flickr)

Nolan Peers (Nolan Peers on flickr)

Jeremy Brooks (Jeremey Brooks on flickr)

(auer1816 on flickr)

"Phillip" (Photog*Phillip on flickr)

Melissa Segal (Melissa Segal on flickr)

Paul Baxter (Mr.Baxter on flickr)

"julián" (Broken Piggy Bank on flickr)

Matt Bucy (aloofdork on flickr)

FORM **NTIA-29** U.S. DEPARTMENT OF COMMERCE
(4-80) NAT'L. TELECOMMUNICATIONS AND INFORMATION ADMINISTRATION

BIBLIOGRAPHIC DATA SHEET

1. PUBLICATION NO. TR-09-459	2. Government Accession No.	3. Recipient's Accession No.

4. TITLE AND SUBTITLE Relationships Between Intelligibility, Speaker Identification, and the Detection of Dramatized Urgency	5. Publication Date Nov. 2008 6. Performing Organization NTIA/ITS.T
7. AUTHOR(S) Andrew A. Catellier and Stephen D. Voran	9. Project/Task/Work Unit No. 6513000-320
8. PERFORMING ORGANIZATION NAME AND ADDRESS Institute for Telecommunication Sciences National Telecommunications & Information Administration U.S. Department of Commerce 325 Broadway Boulder, CO 80305	10. Contract/Grant No.
11. Sponsoring Organization Name and Address National Institute of Standards and Technology/Office of Law Enforcement Standards 100 Bureau Drive, M/S 8102 Gaithersburg, MD 20899-8102	12. Type of Report and Period Covered

14. SUPPLEMENTARY NOTES

15. ABSTRACT

The systems used for public safety speech communications must be intelligible. It is also desirable that they transmit secondary information, such as the attributes of a speaker's voice. This secondary information can allow a user to identify the speaker and his or her emotional state. Testing speech communications systems for the delivery of intelligible speech is common. Testing for human perception of the delivery of this secondary information is less common, though some prior work has been done. Building on this prior work, we describe a set of controlled laboratory listening experiments. These experiments characterize the relationships between speech intelligibility, speaker identification, and the detection of dramatized urgency in a speaker's voice across a range of simulated speech processing conditions. The experiment results indicate that for the speech processing conditions considered here, detection of dramatized urgency is the most robust property, speaker identification is less robust, and speech intelligibility is the least robust.

16. Key Words

human listening tests; intelligible speech; speaker identification; speaker stress detection; speaker urgency detection; speech transmission system; subjective speech quality tests

17. AVAILABILITY STATEMENT □ UNLIMITED.	18. Security Class. (This report) Unclassified	20. Number of pages 35
	19. Security Class. (This page) Unclassified	21. Price:

NTIA FORMAL PUBLICATION SERIES

NTIA MONOGRAPH (MG)

A scholarly, professionally oriented publication dealing with state-of-the-art research or an authoritative treatment of a broad area. Expected to have long-lasting value.

NTIA SPECIAL PUBLICATION (SP)

Conference proceedings, bibliographies, selected speeches, course and instructional materials, directories, and major studies mandated by Congress.

NTIA REPORT (TR)

Important contributions to existing knowledge of less breadth than a monograph, such as results of completed projects and major activities. Subsets of this series include:

NTIA RESTRICTED REPORT (RR)

Contributions that are limited in distribution because of national security classification or Departmental constraints.

NTIA CONTRACTOR REPORT (CR)

Information generated under an NTIA contract or grant, written by the contractor, and considered an important contribution to existing knowledge.

JOINT NTIA/OTHER-AGENCY REPORT (JR)

This report receives both local NTIA and other agency review. Both agencies' logos and report series numbering appear on the cover.

NTIA SOFTWARE & DATA PRODUCTS (SD)

Software such as programs, test data, and sound/video files. This series can be used to transfer technology to U.S. industry.

NTIA HANDBOOK (HB)

Information pertaining to technical procedures, reference and data guides, and formal user's manuals that are expected to be pertinent for a long time.

NTIA TECHNICAL MEMORANDUM (TM)

Technical information typically of less breadth than an NTIA Report. The series includes data, preliminary project results, and information for a specific, limited audience.

For information about NTIA publications, contact the NTIA/ITS Technical Publications Office at 325 Broadway, Boulder, CO, 80305 Tel. (303) 497-3572 or e-mail info@its.bldrdoc.gov.

This report is for sale by the National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22161,Tel. (800) 553-6847.