Speech Intelligibility and Detection of Voice Characteristics

Homeland Security

DHS-TR-PSC-08-05

Department of Homeland Security

Public Safety Communications

Technical Report

This page intentionally left blank.

Homeland Security
Science and Technology

Command, Control and Interoperability: Communication, Interoperability and Compatibility

Defining the Problem

Emergency responders—police officers, fire personnel, emergency medical services-need to share vital voice and data information across disciplines and jurisdictions to successfully respond to day-to-day incidents and large-scale emergencies. Unfortunately, for decades, inadequate and unreliable communications have compromised their ability to perform mission-critical duties. Responders often have difficulty communicating when adjacent agencies are assigned to different radio bands, use incompatible proprietary systems and infrastructure, and lack adequate standard operating procedures and effective multi-jurisdictional, multi-disciplinary governance structures.

OIC Background

The Department of Homeland Security (DHS) established the Office for Interoperability and Compatibility (OIC) in 2004 to strengthen and integrate interoperability and compatibility efforts to improve local, tribal, state, and Federal emergency response and preparedness. Managed by the Science and Technology Directorate, and housed within the Communication, Interoperability and Compatibility thrust area, OIC helps coordinate interoperability efforts across DHS. OIC programs and initiatives address critical interoperability and compatibility issues. Priority areas include communications, equipment, and training.

OIC Programs

OIC programs, which are the majority of Communication, Interoperability and Compatibility programs, address both voice and data interoperability. OIC is creating the capacity for increased levels of interoperability by developing tools, best practices, technologies, and methodologies that emergency response agencies can immediately put into effect. OIC is also improving incident response and recovery by developing tools, technologies, and messaging standards that help emergency responders manage incidents and exchange information in real time.

Practitioner-Driven Approach

OIC is committed to working in partnership with local, tribal, state, and Federal officials to serve critical emergency response needs. OIC's programs are unique in that they advocate a "bottom-up" approach. OIC's practitioner-driven governance structure gains from the valuable input of the emergency response community and from local, tribal, state, and Federal policy makers and leaders.

Long-Term Goals

- Strengthen and integrate homeland security activities related to research and development, testing and evaluation, standards, technical assistance, training, and grant funding.
- Provide a single resource for information about and assistance with voice and data interoperability and compatibility issues.
- Reduce unnecessary duplication in emergency response programs and unneeded spending on interoperability issues.
- Identify and promote interoperability and compatibility best practices in the emergency response arena.

This page intentionally left blank.

Public Safety Communications Technical Report

Speech Intelligibility and Detection of Voice Characteristics

DHS-TR-PSC-08-05
August 2008

Reported for: **The Office for Interoperability and Compatibility**
by NIST/OLES

Homeland Security

This page intentionally left blank.

Publication Notice

Disclaimer

The U.S. Department of Homeland Security's Science and Technology Directorate serves as the primary research and development arm of the Department, using our Nation's scientific and technological resources to provide local, state, and Federal officials with the technology and capabilities to protect the homeland. Managed by the Science and Technology Directorate, the Office for Interoperability and Compatibility (OIC) is assisting in the coordination of interoperability efforts across the Nation.

Certain commercial equipment, materials, and software are sometimes identified to specify technical aspects of the reported procedures and results. In no case does such identification imply recommendations or endorsement by the U.S. Government, its departments, or its agencies; nor does it imply that the equipment, materials, and software identified are the best available for this purpose.

Contact Information

Please send comments or questions to: S&T-C2I@dhs.gov

This page intentionally left blank.

Contents

Abstract

This report describes a laboratory study on the suitability of speech transmission systems. Specifically, public safety first responders listened to and evaluated a large number of recordings of speech transmission systems. The packet loss requirements given in Section 2 of the public safety Statement of Requirements (PS SoR) Volume II [1] are based on the results of this laboratory study.

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, but 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 the design, implementation, analysis and results of 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 wide 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). It is essential that said public safety communication system preserves intelligibility of 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. If transmitted successfully, this secondary information allows the listener to identify the speaker or to confirm the purported identity of the speaker. It also may be possible to identify the emotional state of the speaker. It is generally desirable that a speech communication system transfers this secondary information in addition to providing intelligible speech.

The ability to identify or confirm the identity of a speaker (speaker identification, or 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 ("This is Officer Roberts speaking.") can be avoided. If it is possible to detect that a speaker is not as claimed, this could be very important indeed.

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 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 (e.g., from a neutral tone to a tone of urgency) is detected, the listening public safety officials will certainly want to commit full attention to that specific speaking official to provide support and aid as possible, given the situation.

This report describes the design, implementation, analysis and results of a set of three controlled laboratory listening experiments. An intelligibility experiment finds the word intelligibility (in sentence context)

associated with a set of six different speech processing conditions. A SID experiment characterizes the ability of listeners to identify six different speakers when those speakers are heard after the six different speech processing conditions have been applied. A third experiment characterizes listener detection of dramatized urgency for recordings that have been subjected to those same six speech processing conditions. In the third experiment, listeners attempt 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 characterize 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. Testing for human perception of the delivery of the secondary information is less common. In fact, 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. Significant work has been done on the related topics of automatic recognition of speaker emotions [2] and automatic speech recognition that is invariant to speaker emotions [3].

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” [4]. 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 [5].

Uzdy used two different low-rate vocoders to conduct a SID experiment where listeners were familiar with the speakers [6]. His goal was to determine each vocoders’ 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 [7, 8, 9, 10, 11]. In [7] 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 [12].

3 Speech Recordings

3.1 Intelligibility

There are numerous approaches to testing the intelligibility of speech. In many regards, word intelligibility in a sentence context seems most relevant to public safety communications. 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) [13]. We chose this database because it includes semi-spontaneous speech, repeated utterances of lists of sentences and digits, and some utterances are recorded by multiple speakers.

To ensure that the experiment size was manageable within the limits of human memory (as suggested in [7]), we decided to select three female speakers and three male speakers from the database. After determining the average pitch and voicing strength for each speaker, we looked for male speakers that spoke the same sentences and spanned the full range of pitches found in the database. Additional considerations in selecting 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, respectively. 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 [2, 3, 14, 15].

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 [14, 16]. One portion of the SUSAS database involves a male helicopter pilot recording isolated words in neutral (helicopter on the ground and running) and task (pilot flying helicopter) situations. Another portion includes 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. In the context of this report, however, DU seems more relevant than TIS. Thus, the TIS portion of the experiment is not addressed in the remainder of this report.

3.3.2 Dramatized Urgency

It was important that the experiment include speakers conveying urgency. 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 two words to twenty-one words with a median length of nine 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 (neutral) situation 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.

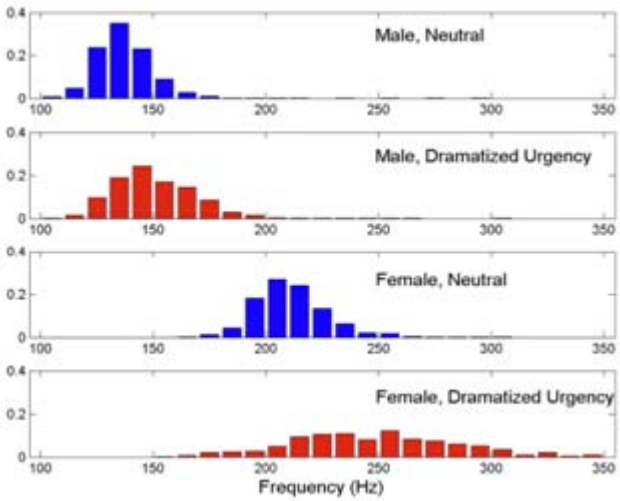
3.3.3 Acoustic Correlates in Dramatized Urgency

We have analyzed the DU recordings and can report several acoustic correlates. The level of DU speech is increased (over 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 [15]. 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 [17]. The level of each recording was then normalized to –26 dB, relative to clipping using tools from [17]. 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. Thus, six speech processing conditions were chosen 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

Table 1: Six conditions used in the speech intelligibility, speaker identification, and detection of dramatized urgency experiments. (Continued)

Condition (C)	Description
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) [18] 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. 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

Evaluation of speech intelligibility, SID, and the detection of DU each require separate laboratory procedures and user interfaces. Note that the speech intelligibility experiment and the detection of DU experiment were conducted in a single multipart laboratory listening session, and thus these experiments used the same listeners. The SID experiment was conducted about 6 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 and in a sound-isolated room where the average background noise level was below 20 dBA. The listening instrument was 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 used 4 times per condition, but only once per listener), for a grand 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 would be considered an outlier in this speech intelligibility task.

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.

All 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 are 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 includes seven different parts. Three are actual experimental test sessions where data is collected, and four are supporting parts that are 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. The listening instrument was 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.

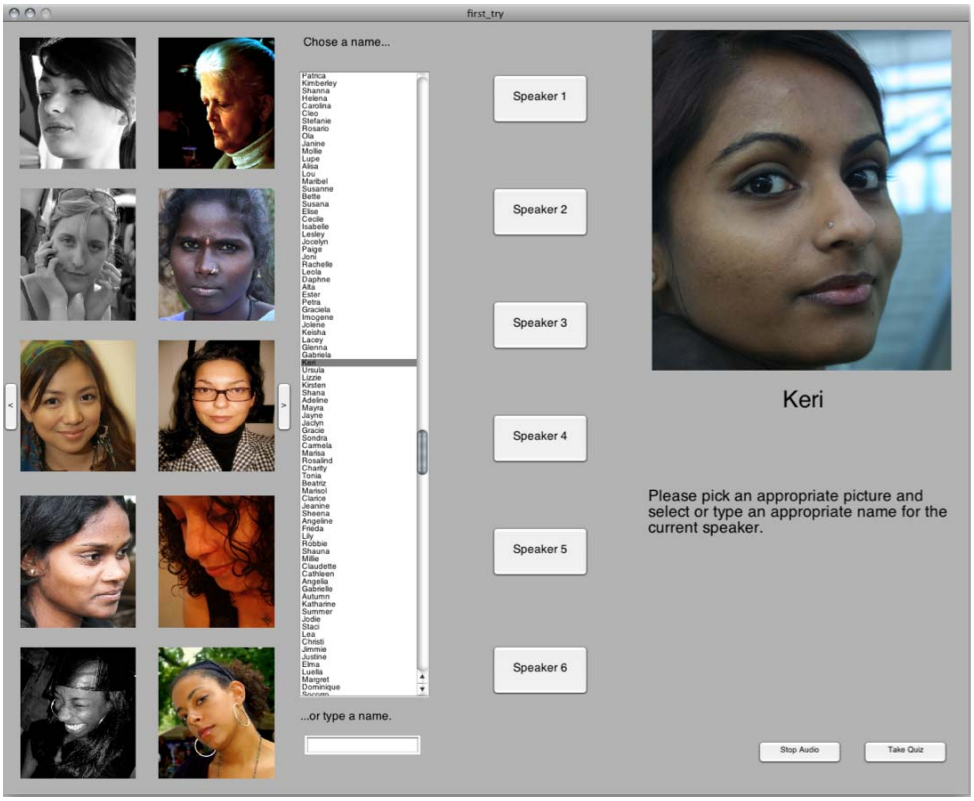
The experiment administrator received many hours of exposure to both undistorted and distorted recordings from the six speakers. After this incidental training, the experiment administrator also served as a listener. As described in Section 6.2.1, his results are not included in the overall experiment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

5.2.2 Preliminaries and Training

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, listeners use a computer. They are instructed to select a name and portrait[1] that is appropriate for the person they hear speaking. The names and portraits are displayed in a window on the computer screen (see Figure 2). After hearing all six speakers and assigning a name and portrait to each, listeners are given a short quiz. A sentence spoken by a random speaker is played back, and the listener is asked to select the appropriate identity from a pool of identities he created. Once an identity has been chosen, the listener confirms his choice. After confirmation, the listener is notified if his selection was accurate. The process iterates through all six speakers, and once finished the listener is allowed to go back to the training session, or move on to the test sessions.

Figure 2: Speaker identification training interface for listeners.



5.2.3 Experimental Test Session 1—Sentences

The first experimental test session uses two sentences from each speaker. Since there are six speakers and six conditions, this results in a total of $2 \times 6 \times 6 = 72$ 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.

1. The portraits shown in Figure 2 were used under Creative Commons license: attributions available on request.

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. Once clicked, the selected identity displayed prominently, 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. After the listener has heard the instructions pertaining to the upcoming session, 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 used in the upcoming session. The listener can spend as much time in this reminder session as is desired, and must 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 four recordings from each speaker. The content of each recording is four spoken digits (e.g., “three six nine eight”). This gives a session with a total of $4 \times 6 \times 6 = 144$ trials. The recordings used 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).

With one exception, all speakers have recorded four unique sets of digits for a total of 15 unique sets of four digits. Pairs of these sets often have two or three digits in common, and indirect SID using content would be extremely difficult, if not impossible.

The third session of the experiment is much like the second session except that the recordings 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). The listening instrument was 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’ 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

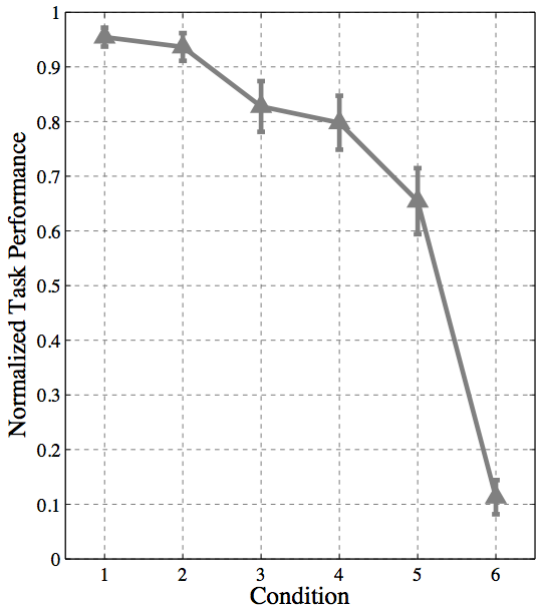
Throughout this section and the next, we report results in terms of normalized task performance (NTP). We introduce this scale because it enables a more direct comparison of the results from the three experiments. On the NTP scale, a value of one indicates perfect information from the listeners in the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

6.1 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 one would be the result. If none of the words were repeated correctly, then an NTP value of zero would be the result.

Figure 3 shows the mean NTP values and 95 percent confidence intervals 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 steadily from 0.95 to 0.11, and this is an NTP drop of 0.84.

Figure 3: NTP mean and 95-percent confidence intervals for word intelligibility in sentence context.



A second version of the intelligibility experiment was completed by six listeners. In this version, listeners could repeat the playing of recordings as they wished. The results are close to those shown for the original experiment in Figure 3. NTP values after the final playing are somewhat greater than those after the initial playing. However, the drop in NTP values (moving from C1 to C6) is 0.79, and this is very close to the value of 0.84 found in the initial experiment. As a result, the final conclusions about relative robustness in 7 are not greatly influenced by the choice of an intelligibility testing approach (single play of each recording versus unlimited playing of each recording).

The increase in NTP associated with multiple plays of the speech recordings ranges from 0.08 to 0.18. The larger increases are associated with cases of medium speech intelligibility where additional plays apparently can help at least some listeners with the task. The smaller increases are associated with cases of very high and very low intelligibility. 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 (C1) to just below two (C6).

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

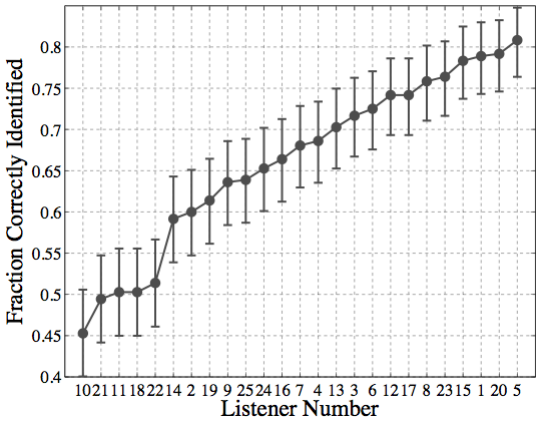
$$\hat{p} = \frac{number\ of\ correct\ identifications}{total\ number\ of\ identifications}. \quad (1)$$

The 95-percent confidence interval for the estimate $\hat{p}$ is calculated as given in [19]. We report $\hat{p}$ as the "Fraction Correctly Identified" in sections 6.2.1 and 6.2.2, and we report the 95-percent confidence interval for the estimate $\hat{p}$ as well.

6.2.1 Listeners

Figure 4 shows the sorted fraction of correct identifications and the associated 95-percent confidence interval for the 25 listeners. The mean fraction of correct identifications over all listeners is .662, and 20 of the 25 listeners have overall correct identification fractions between 0.59 to 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 4 shows that none of these three listeners is an outlier. Thus, all three are retained in our data pool.

Figure 4: Fraction correctly identified by listener and 95-percent confidence intervals.



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 placed close to the fraction-correct mean among all listeners; one of these listeners used a hearing aid. We elected to retain all three of these listeners in our data pool.

Not shown in Figure 4 is the experiment administrator, who received a great deal more training (more than 20 hours on speech distorted under all conditions) and was significantly more accurate with a .98 fraction of correct identifications. This is an indication that additional training can have a positive effect on SID performance, and that the results obtained in this experiment likely form a lower bound for the SID performance to be expected from listeners who have more than a minimal amount of training. 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.

The confusion between the speakers is made precise by a confusion matrix. Table 2 is the confusion matrix for the SID task for these six speakers averaged across all clips, conditions, and listeners. Each row in Table 2 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: Confusion matrix.

	M1	M2	M3	F1	F2	F3
M1	0.67	0.22	0.11	0.00	0.00	0.00
M2	0.15	0.57	0.22	0.01	0.03	0.01
M3	0.12	0.34	0.54	0.00	0.00	0.00
F1	0.00	0.003	0.001	0.65	0.19	0.16
F2	0.00	0.004	0.001	0.17	0.74	0.08
F3	0.001	0.003	0.005	0.07	0.12	0.80

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 .67) and female 1 are close to the same difficulty, and males 2 and 3 (fractions of .57 and .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.

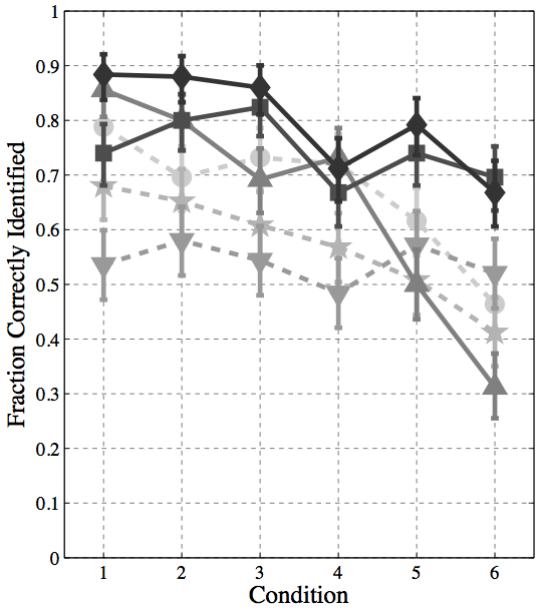
As Schmidt-Nielsen notes, listeners perform the SID task more efficiently with familiar, or distinctive speakers [7]. Our results are consistent with prior research-the two male speakers who had the smallest fraction of correct identifications were also often expressed as perceptually similar by listeners 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.

Listeners received only a small amount of training with these six unfamiliar speaker voices. 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 5. Males 1, 2, and 3 are all shown with dotted lines, and are distinguished by circle, star, and downward-pointing triangle markers, respectively. Females 1, 2, and 3 are all shown with solid lines, and are distinguished by upward-pointing triangle, square, and diamond markers, 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 5: Fraction correctly identified by speaker and 95-percent confidence intervals.



6.2.3 Conditions

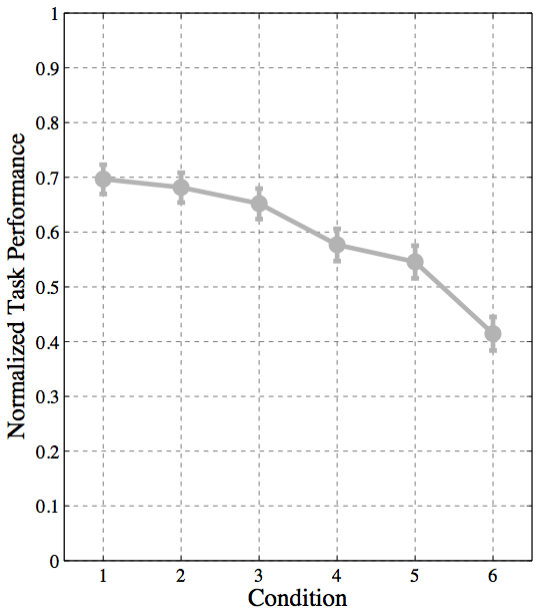
A main goal of this work is to quantify how the listener SID performance is influenced by the six speech processing conditions. Each of the six conditions described in Table 1 was used for a total of 1,500 SID trials. For each condition, these 1,500 trials used the same 60 recorded speech files, and the same 25 listeners as well. This balance allows us to compare SID results for the six conditions directly, as Figure 8 shows. This figure gives results on a NTP scale. On this scale, zero represents no information from listeners, and one represents perfect information from listeners. The transformation from estimated probability of correct identification $\hat{p}$ to NTP is

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

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 zero. On the other hand, perfect SID corresponds to an NTP value of one.

Figure 6: NTP mean and 95-percent confidence intervals for SID task.



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-percent confidence interval for that estimate are calculated as given in (1) and [19]. As with SID, we report detection of DU results in terms of the NTP scale. In this case, that scale is defined by

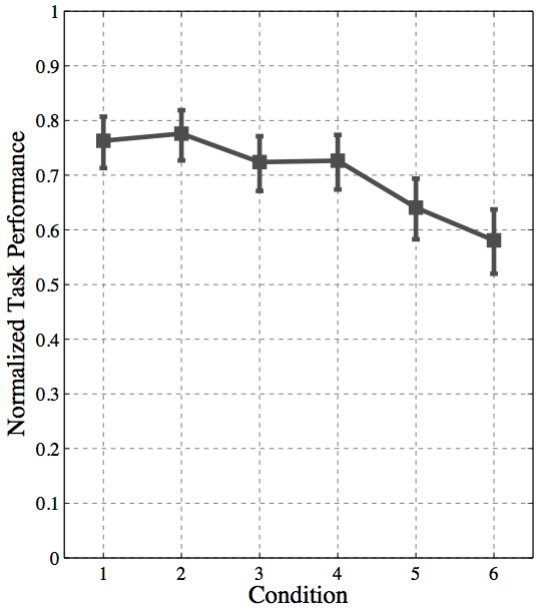
$$\mathrm{NTP} = 2 \times \left(\hat{p} - \frac{1}{2} \right). \tag{3}$$

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 zero.

Figure 7 shows the mean NTP values and 95-percent confidence intervals for each of the six speech processing conditions, after averaging over all listeners and all messages for each condition. Figure 7 shows that as one progresses from C1 to C6, the NTP for detection of DU in messages drops steadily 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

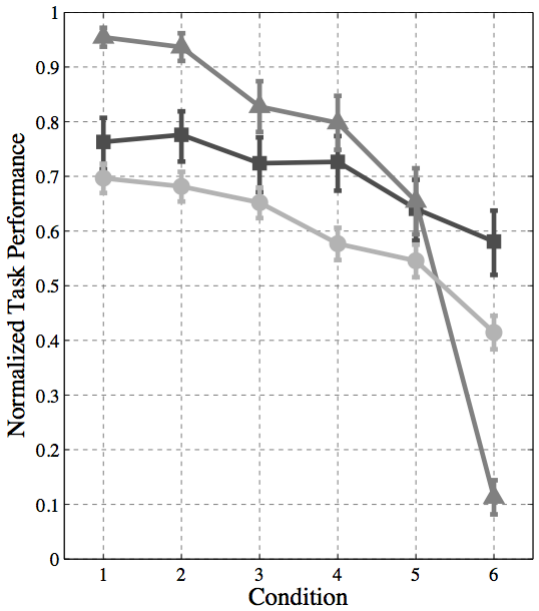
rates generally span the range 0.10 to 0.35. In other words, detection of DU errors are less frequent when speakers are in the neutral state, and more frequent when speakers are in the DU state.

Figure 7: NTP mean and 95-percent confidence intervals for detection of dramatized urgency.



In Figure 8, triangles show NTP mean and 95-percent confidence intervals for word intelligibility; squares show mean and confidence intervals for detection of dramatized urgency; circles show mean and confidence intervals for SID.

Figure 8: NTP mean and 95-percent confidence intervals for word intelligibility, detection of dramatized urgency, and SID.



7 Combined Results and Discussion

This section contains 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 8. All three results generally drop as one progresses from C1 to C6.

Of these three results, the intelligibility results drop most abruptly (from 0.95 to 0.11 for a drop of 0.84), and the detection of DU results drop most gently (from 0.76 to 0.58 for a drop of 0.18). The SID results show a drop that is between the other two (from 0.69 to 0.41 for a drop of 0.28).

If we compare the NTP drop for SID with the NTP drop 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 drop for detection of DU with the NTP drop 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 it has a usable level of word intelligibility (e.g., 80 percent of words intelligible, or NTP of 0.8) then that system will also support good SID performance (e.g., NTP only 0.11 below the best case value), and good detection of DU performance (e.g., NTP only 0.03 below the best case value) as well.

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 consequences 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.

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. Perhaps when immediate attention is critical, these professionals, even if calm by demeanor or by training, could "dramatize urgency" with their voices.

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.

8 References

[1] *Public Safety Statement of Requirements for Communications & Interoperability*, Volume II: Quantitative, Version 1.1, November 2007. Available at: http://www.safecomprogram.gov/SAFECOM/.

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

[3] 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, July 2000.

[4] 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.

[5] 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.

[6] 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, June 1985.

[7] 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.

[8] 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.

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

[10] 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, May 1996, vol. 2, pp. 1149–1152.

[11] 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, November 1998, vol. 2, pp. 221–224.

[12] T. F. Quatieri, *Discrete-Time Speech Signal Processing, Principles and Practice*, Chapter 14, Prentice Hall, Upper Saddle River, New Jersey, 2002.

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

[14] 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*, March 1999, vol. 4, pp. 2079–2082.

[15] 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, November 1996.

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

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

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

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