SPEAKER IDENTIFICATION IN LOW-RATE CODED SPEECH

Andrew Catellier and Stephen Voran

Institute for Telecommunication Sciences
Telecommunications Theory Division
Boulder, Colorado, USA
{acatellier,svoran}@its.bldrdoc.gov

ABSTRACT

While useful speech communication systems must be intelligible, most systems aim to transmit secondary information, such as attributes of a speaker's voice, as well. This secondary information can allow a listener to identify the speaker and his 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 new listening experiment that characterizes the listener identification of six different speakers using six different low-rate digital speech communication systems.

We display these experimental results along with results from our prior work to quantify listener detection of dramatized speaker urgency and word intelligibility in sentence context for the same six speech communication systems. We conclude that the speaker identification task used in this experiment is about three times more robust to communication system degradations than word intelligibility in sentence context.

Index Terms— Speaker identification, speech coding, speech intelligibility, subjective testing

1. INTRODUCTION

Speech communication systems are designed to carry a message from a speaker to a listener. It is essential that said communication system preserves intelligibility of the message. In addition to intelligibility, most systems aim to successfully transmit secondary information, such as attributes of the speaker's voice. If transmitted successfully, this secondary information allows the human listener to identify the speaker (or confirm the purported identity of the speaker) and perhaps even 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 and/or verified implicitly based on transmitted attributes of their voices, the additional overhead associated with explicit identification ("This is Officer Roberts speaking") can be avoided. Similarly, these officials often need to monitor multiple short transmissions with only partial attention while performing other important tasks. When a shift in speaker emotional state is detected (e.g., stress or urgency is perceived) in a specific transmission, the official can then commit full attention to the corresponding speaker.

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. Some studies in the area of 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" [1]. He found that SID can happen with recordings as short as 25 ms, and that high-pass filtering, even with a cutoff frequency as high as 1020 Hz, did not have a significant detriment to SID performance. On the other hand, low-pass filtering at 1020 Hz substantially reduced SID performance. He also found that when the pitch of different speakers was closer, those speakers were more easily confused.

Voiers conducted a fundamental experiment in 1964 [2]. The goal of this work was to identify the main perceptual dimensions of listener SID space. The top four dimensions were identified as "clarity,"

"roughness," "magnitude," and "animation." Later work by Voiers identified eight principle dimensions: "Perceived speech rate," "Pitch magnitude," "Continuity," "Clarity-beauty," "Roughness," "Vocality," "Click-like elements," and "Steadiness" [3].

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 2.4-kbps LPC 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, 7, 8, 9, 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].

Building on this previous work, we have designed, conducted, and analyzed a listening experiment to characterize the ability of listeners to identify six different speakers when those speakers are heard via six different low-rate digital speech communication systems. In the sections that follow, we describe the speech recordings used in the experiment, the six low-rate digital speech communication systems, the experiment design, the software interface developed for the experiment, and the main results obtained. In the results section, we include findings from our previous work using the same speech communication systems for comparison purposes. These results include listener detection of dramatized speaker urgency and word intelligibility in sentence context [12].

2. SPEECH RECORDINGS

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 [6]), 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.

After the desired recordings were extracted from the TSID database, they were resampled to a rate of 8,000 samples per second (from the original sample rate of 16,000 samples per second) using the "PCM filter" option (160 to 3640-Hz bandpass filtering) provided in [14]. The level of each recording was then normalized to -26 dB relative to clipping using tools from [14]. Next, the recordings were passed through software implementations of low-rate digital speech communication systems.

3. COMMUNICATION SYSTEMS

Six experimental conditions were chosen to represent six different communication systems. These are the same conditions used in our previous experiment to characterize listener detection of dramatized speaker urgency and word intelligibility in sentence context [12]. Thus, the description that follows is essentially identical to the description given in [12].

The six systems are summarized in Table 1. C1 involves no additional processing and thus provides a best-case reference point for the SID task. In C4, Modulated Noise Reference Unit (MNRU) [15] software produces multiplicative (speech correlated) noise resulting in an active speech SNR of 6 dB. This does not directly represent any communication system (other than very coarsely quantized PCM or ADPCM) but is included because it is a standardized reference condition that can allow one to build

Condition (C)	Description
C1	Null (no further processing)
C2	IMBE Codec, 7.2-kbps gross 4.4-kbps net
C3	MELP Codec, 1.2-kbps net
C4	MNRU, $Q = 6$ dB SNR
C5	IMBE Codec, 3.6-kbps gross 2.45-kbps net 7-percent BER, random
C6	C5+Packet Impairments+C5
	Packet Impairments: create 60-ms packets, delete 10 percent of packets at random, insert the same number of empty packets at random, and apply PLC to them

Table 1. Six conditions used in the experiment.

relationships to other experiments.

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. The Improved Multiband Excitation Codec (IMBE) running at 4.4 kbps with no bit errors is used for C2. Mixed Excitation Linear Prediction (MELP) at the rate of 1.2 kbps is the coding algorithm used in C3, again with no bit errors. C5 uses IMBE again, but now with 2.45 kbps of speech data and 1.15 kbps of forward error correction. In this condition, 7 percent of the bits are errored using a random bit error distribution.

For C6, three simulated communication systems are concatenated. The first and last are the same as C5 (speech encoding, 7-percent bit errors in the transmission channel, then speech decoding). The middle system consists of packetization of the speech samples into 60-ms packets. Then 10 percent of these packets are deleted at random locations, and an equal number of empty packets are inserted at different random locations. A packet loss concealment (PLC) algorithm is used to extend previous speech samples into these inserted empty packets [16]. This is not a precise model for the behavior of any specific packetized speech system. Rather, it is a packet-based source of impairment designed to provide (when concatenated with two speech coding links) the worst-case reference point for the SID task.

While C2, C3, C5, and C6 are all relevant to low-rate wireless voice communication systems, it is not the primary goal of this experiment to explicitly evaluate these systems. Instead, the primary goal is to evaluate listener performance at SID, and to find relevant relationships among the various results. We view the conditions in Table 1 as a relevant way to generate these results so that they will span a wide range.

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

4. EXPERIMENT DETAILS

The 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.

4.1. 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 click buttons in a window on the computer screen (the window is shown in Figure 1). Each speaker has a unique button, and speakers are assigned a randomly selected button for each different listener. Once a button is clicked, a set of (36) portraits and a list of (1,000+) names appropriate to the speaker's gender appear on the left side of the window. A randomly selected semi-spontaneous speech recording of the appropriate speaker is played back. The listeners are instructed to select a portrait and name (consistent with an internal image triggered by the sound of the voice) by clicking the appropriate portrait and name. Once assigned, the selected portrait and name prominently display on the right side of the window. During this training session, the listener can reassign any previously assigned speaker, and can listen and make assignments for any speaker at any time, in any order. There is no time limit imposed on this or any session. After

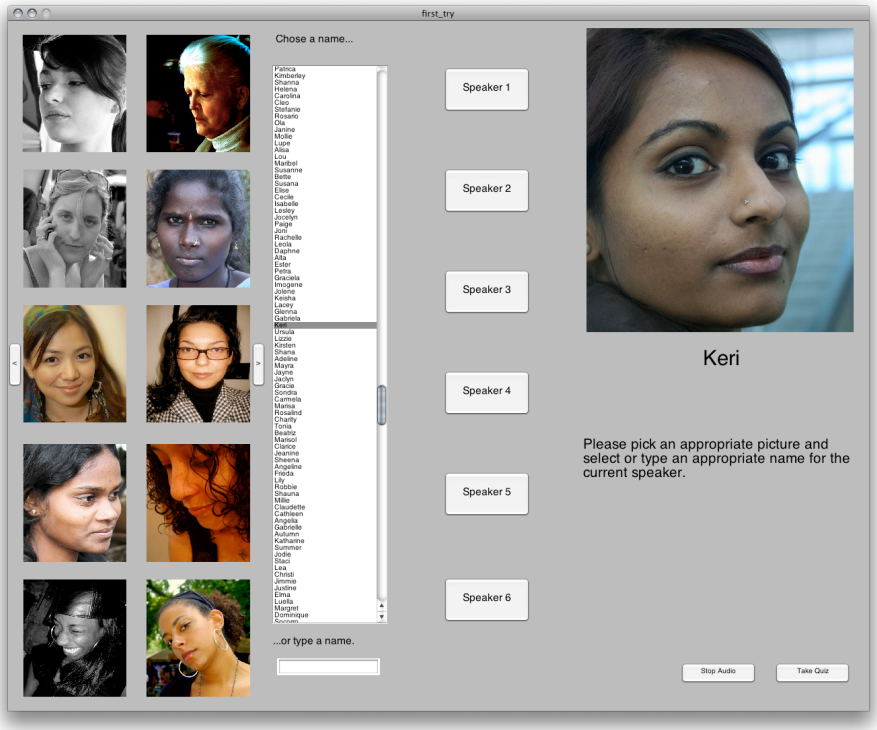

Fig. 1. Screenshot of the training interface. Images used under Creative Commons license: attributions available on request.

a listener has assigned a portrait and name to each speaker, he is instructed to click the button that initiates the quiz session.

The computerized quiz session begins by displaying the six chosen portraits and names in three two-column rows on the left side of the window. Then, a predetermined long sound clip for a speaker (the first in a randomly generated list) is played back. The listener is instructed to identify the speaker of the clip by clicking the portrait that matches the identity of the current speaker.

After the portrait is clicked, it displays prominently (along with the name) on the right side of the window, similar as in the training session. Listeners can click any portrait at any time and are also allowed to replay the sound clip as many times as desired, at any time. The listener is then instructed to click a button labeled "Check Answer" to confirm the choice. If the listener made an incorrect assignment, he is notified that his choice was incorrect, and the identity of the correct choice displays prominently on the right side of the window. If the listener made a correct assignment, the identity remains prominently displayed, and he is notified that his choice was correct. Listeners are allowed to replay the clip as a reminder.

When ready to move on, the listeners click a button labeled "Play Next". A clip from the next speaker in the list is played back, and the process repeats until all six speakers have been identified. Once the quiz session is completed, the listeners are allowed to return to the initial training session, or move on to the first of three experimental test sessions. Listeners are instructed to move on to the experimental test session when they are satisfied with their quiz results. A conceptually similar training system was pioneered in [7].

4.2. 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. 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.

The unique sentence from each speaker does allow the possibility that a clever and ambitious listener might attempt to perform indirect SID by way of content, rather than direct SID by way of acoustic cues. Indirect SID is not desirable in this experiment. One portion of the formal experiment instructions reads, "Please note that you should not try to determine who is speaking by trying to keep track of who has said what—there is no fixed relationship between who is speaking and what is being said."

Our analysis comparing listener performance on the sentence that was the same for all speakers with the sentence that was different for each speaker, showed that the latter had no consistent advantage for correct identification. That is, listeners were not consistently improving their SID performance by using content instead of speaker properties. Further, it was possible for the listeners to hear the same speaker consecutively, possibly leading to secondary training or mistraining effects. However, over all of the experimental test sessions, the same speaker was played twice consecutively only 14 percent of the time, three times consecutively only 2 percent of the time, four times consecutively less than 1 percent of the time, and never five or more times consecutively.

4.3. Experimental Test Sessions 2 and 3—Digits

A short reminder session is provided before 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 $72 + 144 + 144 = 360$.

4.4. 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. The population included three listeners whose first language was not English; Spanish, German and Russian were their native languages. Fourteen listeners reported a scientific or math-based profession, eight reported clerical or other similar desk jobs, and three were IT professionals. 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 below, 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 below, his results are not included in the overall experiment results.

5. RESULTS

As described above, 360 trials were administered to 25 listeners in the main pool. This gives 9,000 data points. In this section we present our analysis of these 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 [17]. In the following, we report $\hat{p}$ as the "Fraction Correctly Identified," and we report the 95-percent confidence interval for the estimate $\hat{p}$ as well.

5.1. Listeners

Figure 2 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 2 shows that none of these three listeners is an outlier. Thus all three are retained in our data pool.

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

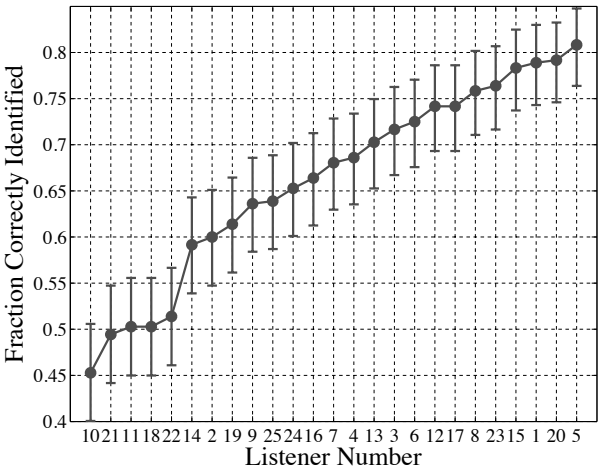

Fig. 2. Fraction correctly identified by listener and 95-percent confidence intervals.

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 on Figure 2 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.

5.2. 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, aver-

aged 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. 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.

	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

Table 2. Confusion Matrix: rows indicate the actual speaker, columns indicate the speaker selected by listeners. "M" indicates male, "F" indicates female. Shaded cells indicate a fraction of correct SID, unshaded cells indicate a fraction of confused SID.

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.

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 3. 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 conditions one, two and four, but is one of the hardest-to-identify speakers when heard over conditions five and six.

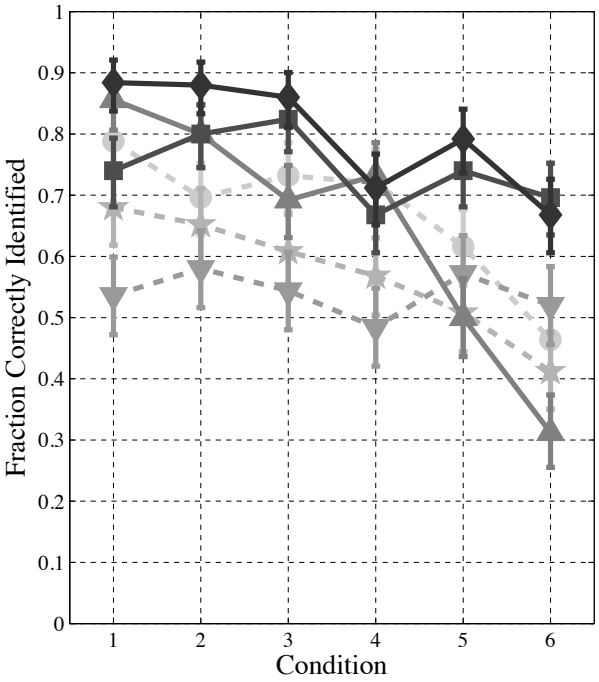

Fig. 3. Fraction correctly identified by speaker and 95-percent confidence intervals. Males 1, 2, and 3 are all shown with dotted lines, and are distinguished by circle, star, and a downward-pointing triangle markers, respectively. Females 1, 2, and 3 are all shown with solid lines, and are distinguished by upward-pointing triangle, square, and diamond markers, respectively.

5.3. Length and Order

Because longer clips provide more speaker information upon which to base a SID decision, intuition might suggest that the SID task would be easier in the case of listening to longer clips rather than shorter clips. However, Figure 4 shows no consistent significant SID performance difference between the three clip lengths (sentence, four spoken digits, two spoken digits) used in this experiment. This may mean that listeners reach their asymptotic performance (versus length) with even the shortest clips (0.7 seconds and 2.5 syllables on average), and would not be at odds with Compton's early work reported in [1].

In our experiment, clip length and presentation order are linked, making it difficult to draw any conclusions about the effect of clip length on SID. Since

the long clips were tested first in the experiment, they may have served as additional training for the listeners before they heard the shorter clips. In fact, in a preliminary experiment design that we tested on several listeners (their results were not used in the data reported in this paper), the first experimental test session used the shortest clips. This ordering seemed to have an extremely detrimental psychological effect: these listeners appeared to lose any confidence they had in their initial training.

While it was difficult to tell if the order of experimental test sessions resulted in additional training, we decided to keep the three different lengths separated due to our belief that the task of identifying a short clip is very different than the task of identifying a long clip. For instance, the longer recordings may contain prosody or personality information that may improve recognition performance. Shorter recordings limit the amount of this information, and listeners may be forced to perform the SID task solely on voice characteristics.

One may argue that our results based on clip length are also confounded with the fact that listeners could replay any clip as many times as they felt necessary. It is true that in public safety applications stimulus length may be fixed and very short due to external circumstances. Since the replay function plays back the exact same recording, there is not any "new" information present that might influence the recognition task. However, by allowing an arbitrary number of replays, and keeping track of the number of replays per recording, we can learn two things: which systems are "hard" for listeners, and about how much extra time is needed for recognition when using any given system. We also believe that not allowing replays may make the task seem harder and ultimately more tiring.

To draw a conclusion about the difficulty of the SID task for different recording lengths, more research must be done.

5.4. Conditions

A main goal of this work is to quantify how the listener SID performance is influenced by communication systems. 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 shown in Figure 5. This figure gives results on a normalized task performance (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

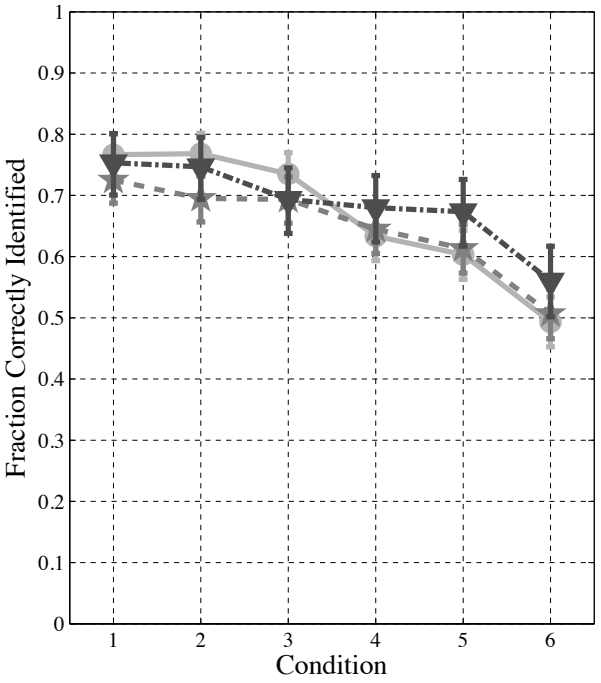

Fig. 4. Fraction correctly identified by length and 95-percent confidence intervals for short clips (shown with circles), medium clips (shown with stars) and long clips (shown with triangles).

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

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

Figure 5 includes two additional results from our prior work [12]. The curve marked with triangles shows the results for word intelligibility in sentence context for the six conditions. Here the NTP value can be interpreted as the percentage of words correctly understood by the listeners.

The curve marked with squares indicates results for the detection of dramatized urgency (DU). This experiment used recordings of speakers reading scripts while verbally dramatizing two different situations: a non-urgent (neutral) situation and a situation requiring an urgent response (DU sit-

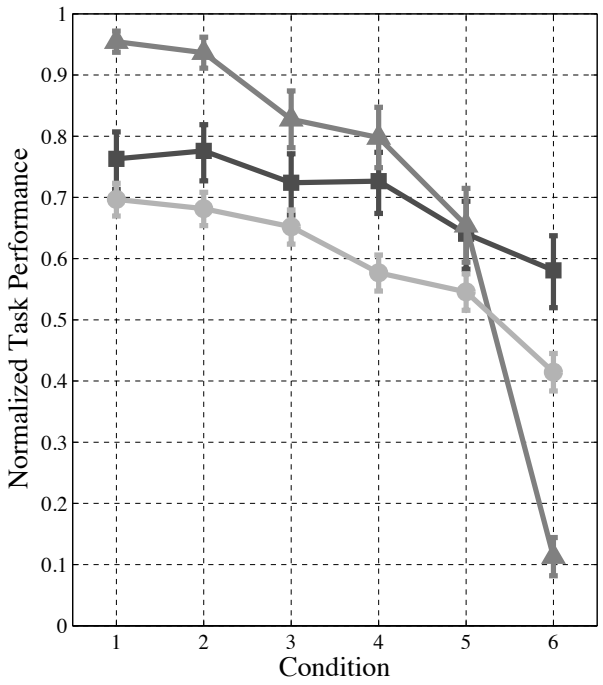

Fig. 5. NTP mean and 95-percent confidence intervals for word intelligibility (shown with triangles), detection of dramatized urgency (shown with squares) and SID (shown with circles).

uation). During recording, 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. Listeners heard the recordings and indicated which category they belong to. Since this was a two-way classification task, a listener making no effort and giving strictly random responses could be correct half the time. Thus for this experiment a correct classification rate of $\frac{1}{2}$ was mapped to an NTP value of zero and perfect classification was mapped to an NTP value of one.

All three curves in Figure 5 show a general trend for decreasing NTP as condition number increases. As one progresses from C1 to C6, the NTP for SID drops from 0.69 to 0.41 (an NTP drop of 0.28) and word intelligibility in sentence context drops from 0.95 to 0.11 (an NTP drop of 0.84). Comparing these two drops in NTP allows us to conclude that for these six conditions, listener identification of the six speakers used in this experiment is about 3 times (0.84/0.28) more robust to communication system degradations than word intelligibility in sentence context. Similarly, the drop in the DU curve shows us that listener detection of DU in this experimental context is about 4.7 times (0.84/0.18) more robust than word intelligibility in sentence context.

6. CONCLUSION

We have designed an experiment to characterize human ability to detect speaker identity. Building on cues from decades of previous research, we implemented an interface that trained listeners to recognize six speakers, using a library of portraits and names as a superset of would-be identity surrogates. We designed a balanced experiment, eliminating or controlling as many variables as possible. Twenty-five listeners identified a healthy fraction of speakers correctly under all conditions, leading to the conclusion that SID, along with detection of DU, are significantly more robust to the distortion present in these communication systems than word intelligibility (3 times and 4.7 times more robust, respectively).

It cannot be concluded that length of stimulus makes a significant contribution to the difficulty of SID in this context. We can conclude, however, that distinctive speakers (despite a relatively small spread of average pitch values) are much easier to identify than similar speakers.

Some speakers are more easily identifiable than others. 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.

It is also important to recognize the fact that hearing disabilities and experience with the language to be recognized did not pose a significant problem in the SID task investigated here.

7. REFERENCES

[1] 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.

[2] W.D. Voiers, "Perceptual bases of speaker identity," *The Journal of the Acoustical Society of America*, vol. 36, no. 6, pp. 1065–1073, 1964.

[3] W.D. Voiers, "Toward the development of practical methods of evaluating speaker recognizability," in *Proc. 1979 IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1979, vol. 4, pp. 793–796.

[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, June 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, September 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, 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 Processsing*, Sydney, November 1998, vol. 2, pp. 221–224.

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

[12] S. Voran, "Listener detection of talker stress in low-rate coded speech," in *Proc. 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, March 2008, pp. 4813–4816.

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

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

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

[16] ITU-T Recommendation G.711, Appendix I, A high quality low-complexity algorithm for packet loss concealment with G.711, Geneva, 1999.

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