LISTENER DETECTION OF TALKER STRESS IN LOW-RATE CODED SPEECH

Stephen Voran

Institute for Telecommunication Sciences, Boulder, Colorado, USA, svoran@its.bldrdoc.gov

ABSTRACT

We describe an experiment where listeners were asked to detect two specific forms of stress in talkers' recorded voices heard via six different simulated communication systems. Both task-induced stress and dramatized urgency were used. Communication systems included low-rate digital speech coding combined with bit errors, packet loss, and packet loss concealment. Twenty-four listeners participated in a total of 11,520 detection trials. A parallel investigation of word intelligibility in sentence context used 576 trials. Intelligibility results showed wide variance due to communication system and stress detection results showed less variance. More specifically, we found that listener detection of dramatized talker urgency was 4.7 times more robust to communication system degradations than word intelligibility in sentence context.

Index Terms**—**Speech coding, speech intelligibility, stress detection, subjective testing, talker stress

1. INTRODUCTION

Speech communication systems can carry information about a talker's emotional state via characteristics of the talker's voice. This capability provides the listener with a sense of realism, but it can also be of vital importance in situations where the listener monitors multiple short transmissions with only partial attention while performing other important tasks (e.g., public safety officials in a field environment). When some form of stress or urgency is perceived in a specific transmission, the listener can then commit full attention to the corresponding talker.

These observations motivate us to consider the ability of a speech communication system to communicate a talker's emotional state as represented in a talker's voice. We have designed, conducted, and analyzed a listening experiment to characterize this ability for two specific simple cases: listener detection of task-induced stress and listener detection of dramatized urgency. Six different communication systems connect talker and listener. For each system, the experiment also produced corresponding intelligibility results for comparison purposes. In the following we describe the speech recordings used and the issues considered in their selection or creation. Next we characterize the recordings, describe the six communication systems simulated in the experiment, and outline the experiment environment and procedures. Finally we report the results obtained and draw conclusions.

2. SPEECH RECORDINGS

The term "stress" is subjective and covers a wide range of circumstances and resulting speech signals. For speech signals, objective refinement of the term "stress" and quantification of stressor levels is enabled through the use of known acoustic correlates. These include changes in level, tempo, pitch and formants [1]-[4].

Our experiment uses speech recordings from two specific scenarios that are outlined in this section. We apply the labels "task-induced stress" (TIS) and "dramatized urgency" (DU) to these two specific groups of recordings and to the corresponding experimental results.

We are not aware of any previous efforts to characterize communication systems' ability to preserve detection of talker stress. However, significant work has been done on automatic recognition of talker emotions [1] and automatic speech recognition that is invariant to talker emotions [2]. Efforts in this second area include the Speech under Simulated and Actual Stress (SUSAS) recorded speech database [3],[5] and we were able to extract portions of this database for the TIS portion of our experiment.

One portion of the SUSAS database we extracted involves a male helicopter pilot recording isolated words in neutral (helicopter on the ground and running) and task (pilot flying helicopter) situations. The second portion we extracted includes one male and one female talker recording isolated words in neutral (no task) and computer-graphics based "dual tracking" task situations. We took care to extract only portions of the database that could form pairs that were nearly free of variations in background noises and recording imperfections. When comparing the task recordings with the neutral recordings we perceived only a minor sense of distraction.

It was important for the experiment to include talkers conveying urgency. It would not be ethical to subject talkers to events (e.g., physical dangers) that could create a true sense of urgency. Recording talkers confronted by naturally occurring urgency-inducing events was not a practical option for us at this time but researchers might consider this 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 talker recorded the DU scripts. We used studio-grade digital recording equipment and a quiet recording room with average noise level below 20 dBA. Each talker 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 talkers should dramatize urgency.

Using the same talkers, equipment, and room we created recordings to support testing of open-set word intelligibility in sentence context. We selected and recorded 20 sentences from current issues of *The Wall Street Journal* and *The New York Times* (WSJ/NYT). 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"). Table 1 summarizes the various recordings, sub-experiment (SE) numbers, and listener tasks.

3. 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 talker and 8.0 dB for the female talker. (Note however 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 talkers responded oppositely in terms of tempo. The male talker 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 talker increased from 134 Hz (neutral) to 148 Hz (DU) while the standard deviation increased from 21 to 23 Hz. For the female talker the mean pitch increased from 211 to 249 Hz and standard deviation increased from 18 to 38 Hz. All four of these results can be seen in the pitch histograms in Figure 1. We also observed changes in formant structure for both talkers.

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 [4].

SE	Speech Recordings	Talkers	Listener Task
1	4 Words Dramatized Urgency	1F, 1M	Rate talker stress or urgency: low or high
2	4 Words, SUSAS Helicopter Pilot	1M	Rate talker stress or urgency: low or high
3	4 Words, SUSAS Dual Tracking Task	1F, 1M	Rate talker stress or urgency: low or high
4	24 Sentences WJS/NYT	1F, 1M	Repeat sentence heard
5	20 Messages Dramatized Urgency	1F, 1M	Rate talker stress or urgency: low or high

Table 1. Summary of speech recordings and listener tasks for each sub experiment (SE).

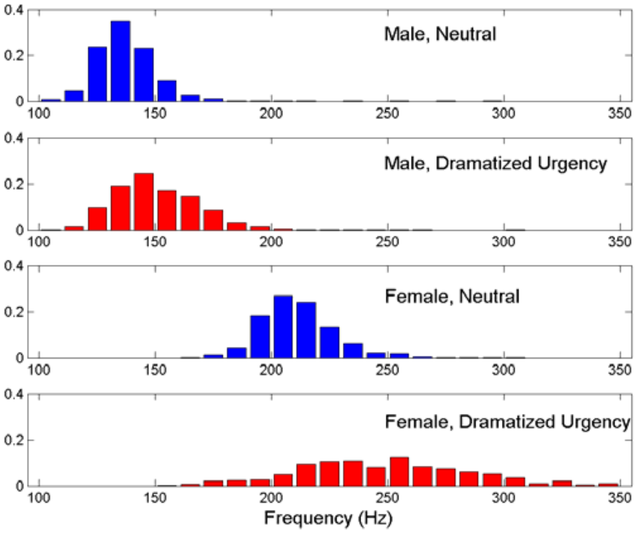

Fig. 1. Pitch histograms for four cases as labeled.

Whether or not DU is a good surrogate for true urgency will likely depend on numerous factors including individual talkers' physical and psychological characteristics and the details of the urgent situation.

4. COMMUNICATION SYSTEMS

Six experimental conditions were chosen to represent six different communication systems. These are summarized in Table 2. To produce a condition, speech recordings were first converted from 48,000 to 8000 sample/sec and then normalized to an active speech level of −26 dB relative to clipping using ITU-T standardized speech processing software tools found in Recommendation G.191. After any relevant software processing for the given condition, the level normalization was applied again as a final processing step. Condition 1 involves no further processing and thus provides a high-quality reference point.

In Condition 4, Modulated Noise Reference Unit (MNRU) software adds multiplicative (speech correlated) noise resulting in an active speech SNR of 6 dB. This does not directly represent any communication system (other

Cond.	Description	Parameters
1	Null	
2	Improved Multiband Excitation Codec	7.2 kb/s, 0% BER
3	Mixed Excitation Linear Prediction Codec	1.2 kb/s, 0% BER
4	MNRU	Q=6 dB (SNR)
5	Improved Multiband Excitation Codec	3.6 kb/s, 7% BER
6	Improved Multiband Excitation Codec	3.6 kb/s, 7% BER
	+ Packetization	60 ms packets, 10% packet impairment rate
	+ Improved Multiband Excitation Codec	3.6 kb/s, 7% BER

Table 2. Six conditions included in the experiment.

than coarsely quantized PCM or ADPCM) but is included because it is a standardized reference condition that can allow one to build 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 bit-rates reported for Conditions 2, 5, and 6 include forward error correction. All bit error patterns are random (uncorrelated) and different for each instance.

In Condition 6 three simulated communication systems are concatenated. The first and last are the same as Condition 5 (speech encoding, 7% bit errors in transmission channel, then speech decoding). The middle system consists of packetization of the speech samples into 60 ms packets, then 10% of these packets are deleted (random packet loss) and an equal number of empty packets are inserted at different random locations. A packet loss concealment algorithm is used to extend previous speech samples into these inserted empty packets.

Note that while Conditions 2, 3, 5, and 6 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 detection of TIS or DU as well as word intelligibility, and to find relationships among the results. We view the conditions in Table 2 as a relevant way to generate these results so that they will span a wide range.

5. LISTENING EXPERIMENT

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 at any time.

Listeners participated in two practice sessions and six actual sessions (SE 5 was divided into 2 sessions). The total time required was typically around one hour. Responses were collected using a GUI on a PDA supported by a wireless LAN connection.

In SEs 1, 2, 3, and 5 (the detection SEs) listeners heard a recording and responded to the prompt “Please select the talker's stress or urgency level.” Response options in each of these binary forced-choice trials were “Low” (the correct answer for neutral recordings) and “High” (the correct answer for TIS and DU recordings). Listeners could respond at any time once a recording had started to play, and could restart the playback at any time. In this manner, each listener could proceed at an individualized pace through the SEs. In SE 1, for example, each listener heard 96 trials (2 talkers × 2 talker states × 4 words × 6 conditions). Overall, each of 24 listeners heard 480 trials for a grand total of 11,520 detection trials in the experiment.

In SE 4 (the intelligibility SE) 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. Each listener 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, for a grand total of 576 trials. Within each SE each listener heard the recordings in a different random order.

6. RESULTS AND CONCLUSIONS

For each trial in a detection experiment three outcomes are possible: correct detection, false alarm (low stress or urgency reported as high), and miss (high reported as low). Figure 2 summarizes the fraction correct (across all talkers, words or messages, conditions, and listeners) for the four detection SEs. Given the binary nature of the data (correct or not correct), the 95% confidence intervals shown in Figure 2 reflect the confidence in the estimated mean of the binomial distribution [6]. The figure shows that the detection of DU in words (SE 1) or messages (SE 5) is significantly easier than detection of TIS in words. Note that since this is a binary response situation, if listeners consistently reply at random, the resulting fraction correct would be 0.5.

Figure 3 provides results for SE 5 by condition. It shows several significant differences in listeners' ability to detect DU in messages as a function of condition, and also shows a general trend for decreasing detection as condition number increases. Figure 3 also summarizes the results of SE 4 (word intelligibility) over all talkers, listeners, and sentences

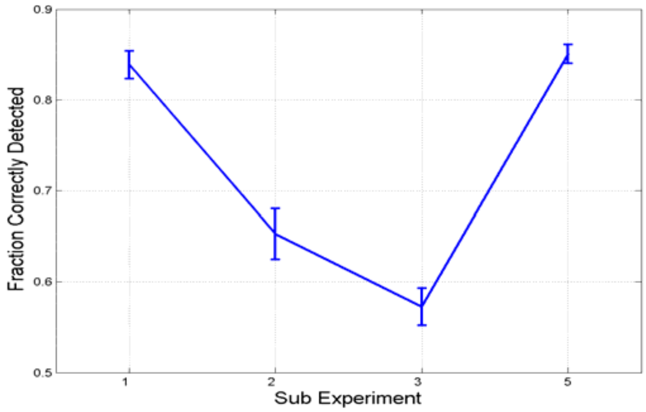

Fig. 2. Fraction of correct responses in four SEs.

for each condition. Again, significant differences are seen, and the general trend is for decreasing intelligibility as condition number increases.

To allow meaningful comparison of results from SEs 4 and 5, Figure 3 uses a normalized task performance (NTP) scale. On this scale zero represents no information from listeners and one represents perfect information from listeners. In SE 4 (intelligibility), the NTP value is simply the fraction of words correctly identified. For SE 5 (detection) NTP = 2·(fraction correct detections) − 1, so that the range [0.5, 1] is mapped onto the NTP range [0, 1].

Figure 3 shows that as one progresses from Condition 1 to Condition 6, the NTP for detection of DU in messages drops from 0.76 to 0.58 (an NTP drop of 0.18) 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 6 conditions, listener detection of DU is about 4.7 times (0.84/0.18) more robust to communication system degradations than word intelligibility in sentence context.

NTP results for SE 1 (DU with words) show no significant differences due to conditions. Detection is fairly easy in this case, and specific conditions do not make it significantly easier or harder. NTP results for SE 2 show no significant differences and NTP results for SE 3 show two barely significant differences. Detection is rather difficult in SEs 2 and 3 (TIS with words) and specific conditions do not make it much easier or harder.

We found that across the SEs and conditions, the false alarm rate tends to be lower (0.05 to 0.10) and the miss rate tends to be higher (0.10 to 0.35). In other words, listener detection errors are less frequent when talkers are in the neutral state.

We conclude that in the context of the six conditions used, TIS in words is difficult to detect and nearly invariant to conditions. DU in words is easier to detect and also invariant to conditions. Detection of DU in messages is similar to DU in words on average, but does show significant variation due to the six conditions. However, the word intelligibility results show 4.7 times more sensitivity to communication system degradations than the DU message detection results.

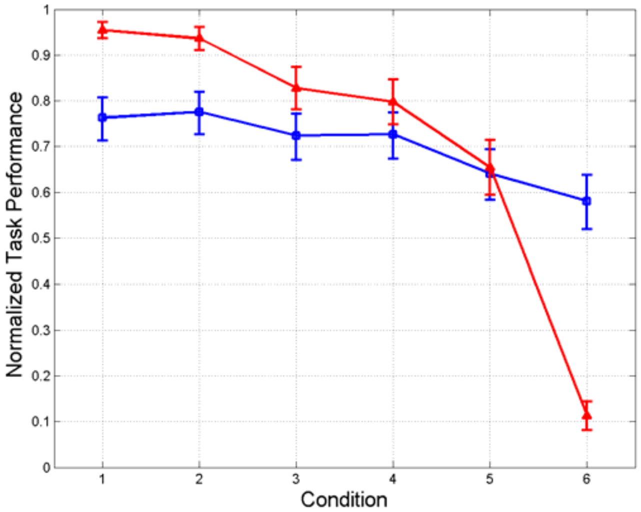

Fig. 3. NTP mean and 95% confidence intervals for SE 4 (word intelligibility, shown with triangles) and SE 5 (detection of DU in messages, shown with squares.)

Based on Figure 3, if a communication system (that is well represented by the six used here) has a usable level of word intelligibility (e.g., 80%) then we would also expect that it would permit listener detection of DU that is only slightly below the maximum possible level (e.g., NTP of 0.73, down from the maximum possible of 0.76). We close with a final note on the potential relevance of DU. Perhaps when necessary, a talker (even if calm by demeanor or by professional training) in need of immediate attention could dramatize urgency in order to achieve the talker detection results described here.

7. REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, J. Taylor, "Emotion Recognition in Human-Computer Interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[2] S. Bou-Ghazale and J. Hansen, "A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech under Stress," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 4, pp. 429-442, Jul. 2000.

[3] H. Steeneken and J. Hansen, "Speech under Stress Conditions: Overview of the Effect on Speech Production and on System Performance," *Proc. IEEE ICASSP 1999,* vol. 4, pp. 2079-2082, Mar. 1999, Phoenix.

[4] R. Ruiz, E. Absil, B. Harmegnies, C. Legros, and D. Poch, "Time- and Spectrum-Related Variabilities in Stressed Speech under Laboratory and Real Conditions," *Speech Communication*, vol. 20, no. 1-2, pp. 111-129, Nov. 1996.

[5] SUSAS Database is available at www.ldc.upenn.edu.

[6] N. Johnson, S. Kotz, and A. Kemp, *Univariate Discrete Distributions, Second Edition*, p. 129, Wiley, New York, 1992.