GETTING STARTED WITH SUSAS:
A SPEECH UNDER SIMULATED AND ACTUAL STRESS DATABASE

John H.L. Hansen and Sahar E. Bou-Ghazale
Robust Speech Processing Laboratory
Duke University, Department of Electrical Engineering
Box 90291, Durham, North Carolina 27708-0291, U.S.A.
http://www.ee.duke.edu/Research/Speech

ABSTRACT

It is well known that the introduction of acoustic background distortion and the variability resulting from environmentally induced stress causes speech recognition algorithms to fail. In this paper, we discuss SUSAS: a speech database collected for analysis and algorithm formulation of speech recognition in noise and stress. The SUSAS database refers to Speech Under Simulated and Actual Stress, and is intended to be employed in the study of how speech production and recognition varies when speaking during stressed conditions. This paper will discuss (i) the formulation of the SUSAS database, (ii) baseline speech recognition using SUSAS data, and (iii) previous research studies which have used the SUSAS data base. The motivation for this paper is to familiarize the speech community with SUSAS, which was released April 1997 on CD-ROM through the NATO RSG.10.

1. INTRODUCTION:
Why Speech Recognizers Break

The issue of robustness in speech recognition can take on a broad range of problems. A speech recognizer may be robust in one environment and inappropriate for another. In Fig. 1, a general speech recognition scenario is presented which considers a variety of speech signal distortions. For this scenario, we assume that a speaker is exposed to some adverse environment, where ambient noise is present and a stress induced task is required. The adverse environment could be a noisy automobile where cellular communication is used, noisy helicopter or aircraft cockpits, noisy factory environments, and others. Since the user task could be demanding, the speaker is required to divert a measured level of cognitive processing, leaving formulation of speech for recognition as a secondary task.

Workload task stress has been shown to significantly impact recognition performance [1, 4, 6, 9, 11, 12, 7]. Since background noise is present, the speaker will experience the Lombard effect [10, 5, 8]; a condition where speech production is altered in an effort to communicate more effectively across a noisy environment. In addition, a speaker may also experience situational stress, (i.e., anger, fear, other emotional effects), which will alter the manner in which speech is produced. If we assume $s(n)$ to represent a Normal, noise-free speech signal, then the acoustic signal at the microphone will include distortion due to stress, Lombard effect, and additive noise.

![Diagram of speech recognition system](http://www.isca-speech.org/archive)

Figure 1: Types of distortion which can be addressed for robust speech recognition.

Other distortions such as microphone mismatch, cellular/telephone channels, or voice coding effects can also degrade the speech signal. Therefore, the Neutral speech signal $s(n)$, is transformed into the degraded signal $y(n)$.

$$y(n) = \left\{ \begin{array}{ll} [s(n) \text{STRESS WORKLOAD TASK LOMBARD EFFECT } d_1(n)] + d_2(n) & \text{if STRESS} \\ [s(n) \text{STRESS WORKLOAD TASK }] + d_1(n) & \text{otherwise} \end{array} \right.$$  

Having outlined the basic structure of a degraded speech signal under stress and noise, we now turn to the formulation of the SUSAS database. This database was designed for analysis and modeling of speech under stress in order to improve the robustness of speech recognition algorithms[4]. Sec. 2 presents an overview of the structure of the SUSAS database. A discussion of the stress styles and workload tasks are also included. Sec. 3 presents baseline speech recognition results to be used for comparison by users of the speech corpus. Finally, Sec. 4 will present a brief summary of previous research studies which have used SUSAS data.

2. STRUCTURE OF SUSAS

SUSAS represents a comprehensive speech under stress database which was formulated and collected by Hansen[4]. The database is partitioned into five domains, encompassing a wide variety of stresses and emotions. A total of 32 speakers (13 female, 19 male), with ages ranging from 22 to 76 were employed to generate in excess of 16,000 utterances. The five stress domains include: i) talking styles\(^1\) (slow, fast, soft, loud, angry, dear, question), ii) single tracking task or speech produced in noise (Lombard effect), iii) dual tracking computer response task, iv) actual subject motion-fear tasks (G-force, Lombard effect, noise, fear), v) psychiatric analysis data (speech under depression, fear, anxiety). A common highly

---

\(^1\) Approximately half of the SUSAS data base consists of style data donated by MIT Lincoln Laboratory (Lippmann, et al. [9], Chen [1], Hansen [4]).
confusable vocabulary set of 35 aircraft communication words make up the database. Simulated speech under stress data consists of data from ten stressed styles (talking styles, single tracking task and Lombard effect domains); while actual speech under stress data consists of speech produced while performing either (i) dual-tracking workload computer tasks, or (ii) subject motion-fear tasks (subjects in roller-coaster rides). Additional speech was also added from pilots in Apache helicopter flight conditions. Two of the domains employ computer response tasks, and one domain employs amusement park roller-coaster rides as speaker tasks. The four domains available in the present release of SUSAS consists of a 35 aircraft communication word vocabulary which is summarized in Table 1. Subsets include (go, hello, oh, no), (six, fix), (white, wide, point), (degree, three, thirty, freeze), and (eight, eighty, gain, change).

D1: Speaking Styles. The first SUSAS Database domain involves speech under various speaking styles. The data in this portion was donated by Lincoln Laboratory [11], and downsampled from 16 to 8 kHz. This portion of SUSAS contains utterances from nine male speakers under eight speaking styles (normal, slow, fast, soft, loud, question, clear, angry).

D2: Single Tracking Task. This SUSAS domain consists of speech data produced while performing a single response computer tracking task. The task is based on the response of a marginally stable, single-pole system. This domain also includes a small portion of speech produced in noise, simulating the Lombard effect [10]. Pink noise was presented binaurally at an overall level of 85 dB SPL.

D3: Dual Tracking Task. The third domain consists of speech produced while performing dual (compensatory and acquisition) tracking tasks which were developed for the USAF School of Aerospace Medicine [2]. A weighted RMS error was found for each task individually, and combined. RMS task error ranged from 31-60% for the speakers, and the high workload case was significantly more taxing than the moderate case (a 9.5% difference).

D4: Actual Motion-Fear Task. This SUSAS domain consists of speech produced during the completion of two types of subject motion-fear tasks. In order to attempt a simulation of the sudden change in altitude/direction experienced in an aircraft cockpit, two types of motion tasks were considered. These tasks were chosen because they required no training, yet generated the type of stress (fear or anxiety) which might be experienced in an emergency situation. Two rides from an amusement park were chosen as suitable, the Scream Machine and Free Fall. Both possess large vertical drops (as much as 30 meters) and sudden horizontal turns. The Free Fall ride lasts for about 60 seconds, with the free fall portion comprising about 10 seconds. Four seated passengers are strapped in an upright seated position into a car which is raised vertically to approximately 40 meters. The car is moved forward where it pauses for several seconds before being released. It drops vertically downward for about 30 meters, before rolling onto a horizontal portion of the track for deceleration. During the free fall portion, talkers repeated several pre-chosen words from the 35 word list. Speech was recorded using a high quality head-mounted microphone and cassette recorder unit strapped to the talker’s body.

The second motion task considered was the Scream Machine. The 90 second ride consists of large vertical movements with small amounts of lateral movement during calm periods between drops. Each speaker (3 female, 4 male) repeated words from the 35-word vocabulary and performed the task twice. The chosen subjects were all native Americans with no apparent speech deficiencies. A total of 1642 utterances were collected from speech under task stress. In each subject motion task, at least four factors contributed to the type of speech recorded: g-force variation, background noise, Lombard effect, fear and/or anxiety.

Four additional male speakers were later added which included pilots flying missions in Apache helicopters. These include two pilots operating their (i) helicopter under normal baseline flight conditions on the ground, and (ii) while flying moderate g-force maneuvers while speaking the 35-word vocabulary. An additional set of recordings were included from two other Apache helicopter pilots flying a night mission into the Raleigh/Durham Airport while running low or out-of-fuel.

D5: Psychiatric. The last domain of the database includes speech from patients undergoing psychiatric analysis in a natural doctor-patient environment. Recordings from eight patients (six female, two male) were obtained. Based on informal listening sessions, the overriding emotion present was mild to severe depression. In some cases, brief passages of fear, anxiety, and/or anger were also identified. At the time of the CD-ROM release, we decided not to include this data until we could determine if permission was needed to release patient speech.

3. Baseline Performance
To illustrate the problem of speech recognition in stress and noise, a baseline speaker-dependent, 5-state, discrete-observation HMM speech recognizer

<table>
<thead>
<tr>
<th>35-Word SUSAS Vocabulary Set</th>
<th>Models Trained With</th>
<th>Models Tested With</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral Speech</td>
<td>90.26%</td>
<td>26.67%</td>
</tr>
</tbody>
</table>

Table 4: Speaker-independent recognition results for a neutral trained model when tested with neutral and actual motion-fear stressed speech.
Str essful Sp e e ch R e c o gnition /#28Sp e aker/-Indep ende nt/, Continuous/-density HMM/#29

So L A C Q C/5/0 C/7/0 L om

A vg/1/0

.: the /8/8k ens p er w ord /#28/8 tok ens p er w ord

Stressful Speech Styles Key:
F #7B fast A #7B angry Lom #7B Lombard e/#0Bect noise condition
Sl #7B slo w L #7B loud Q #7B question C/7/0 /#7B High Load Computer T ask Condition

Table 2: Recognition performance of speaker-dependent neutral trained discrete density HMM tested with neutral and stressed type speech in noise free and noisy conditions. @Additive white Gaussian noise, SNR = +30dB


4. SUMMARY of KEY SUSAS REFERENCES
In this section, we present a brief summary of previous research studies which have used data from the SUSAS Speech Under Stress database. A more complete summary is included on the CD-ROM.

4.1 Stressed Speech Analysis

4.2 Stressed Speech Synthesis
The following papers have considered methods for imparting stress onto neutral input speech.

| Condition   | Stressful Speech Recognition (Speaker-Dependent, Discrete-Density HMM) | N | C | C50 | C70 | Lom | So | Q | L | S | F | A |
|-------------|-----------------------------------------------------------------------|---|---|-----|-----|-----|----|---|---|---|---|---|---|
| Noise-free  |                                                                       | 88.3% | 60% | 66% | 45% | 20% | 68% | 10% | 65% | 63% | 63% | 31.5% | 15.35 |
| Noisy       |                                                                       | 49% | 45% | 38% | 33% | 18% | 15% | 40% | 28% | 35% | 33% | 28% | 30.3% | 9.12 |

**Table 3:** Recognition performance of speaker-independent neutral trained continuous density (2 mixtures per state) HMM models tested with neutral and stressed type speech in noise-free conditions.

(VQ-HMM) was employed on noise-free and noisy stressed speech from SUSAS (Table 2). The speaker-dependent open baseline evaluations show that when stress is present, recognition rates decrease significantly (rates for Lombard effect, loud, and angry speech range from 20-63%, with a neutral rate of 88.3%). When white Gaussian noise is introduced, noisy stressed speech rates varied, with an average rate of 90.5% (i.e., a 58% decrease from the 88.3% neutral rate). Recognition performance also varies considerably across stressed speaking conditions as reflected in the large standard deviation in rate of recognition. (SLDevI= 15.35, 9.12 for noise-free and noisy stressed conditions).

In addition, the baseline recognition performance of a speaker-independent continuous density HMM employing mel-frequency cepstral parameters and time derivatives is also reported (Table 3). The data is parameterized using 12 mel-frequency cepstral coefficients. A 25 msec Hamming window is used, and a first order preemphasis is applied to the data using a coefficient of 0.97. Cepstral mean normalization is performed on the cepstral parameters to compensate for long-term spectral effects. In addition, the cepstral parameters are re-scaled using a cepstral lifter. Delta and acceleration coefficients were also appended to the MFCC coefficients since the performance of a speech recognition system can be greatly enhanced by adding these time derivative features [7]. A total of 35 tokens were used to train each HMM word recognizer. Using 9 male speaker population, a round robin training and testing scenario is employed in order to test all speakers. Hence, a total of 64 tokens per word (8 tokens per word x 8 speakers) per style are employed for training a 2-mixture continuous density 5-state left-to-right HMM model. A total of 2240 tokens per style are employed for testing the speaker-independent recognizer. A 96% recognition accuracy is achieved with models trained and tested with neutral speech. Under stress, the performance varies considerably (73.5% when tested with angry speech).

Baseline recognition performance using the fourth domain of SUSAS are reported in Table 4. A total of 35 tokens were used to train each HMM word recognizer. The speaker population employed in training the neutral models consists of 9 male speakers, while four different male speakers are employed for testing. For each word model, a total of 63 neutral tokens (7 tokens per word x 9 speakers) are employed for training a 2-mixture continuous density 5-state left-to-right HMM model. The same parameter set described previously are used for training and recognition. A total of 575 tokens are employed for testing the neutral trained speaker-independent recognizer. The recognition accuracy of the models trained and tested with neutral speech is 90.26%. Models trained with neutral speech, and tested with actual motion/fear stressed speech achieves a 26.67% recognition accuracy. Hence, the recognition error is 63.59%.

| Models Trained with | Stressful Speech Recognition (Speaker-Independent, Continuous-Density HMM) | N | C | C50 | C70 | Lom | So | Q | L | S | F | A |
|---------------------|-----------------------------------------------------------------------|---|---|-----|-----|-----|----|---|---|---|---|---|---|
| Neutral Speech      |                                                                       | 90% | 95.6% | 95.4% | 95.3% | 90.6% | 90% | 85.9% | 83.6% | 83% | 79.8% | 73.5% |
The following papers have considered methods for detection or classification of speech under stress.


4.3 Stressed Speech Classification

The following papers have considered methods for detection or classification of speech under stress.


5. CONCLUSION

In this paper, we have presented an overview of the SUSAS Speech Under Simulated and Actual Stress Database. A number of baseline recognition evaluations were also presented. The intent of this speech corpus is to assist researchers in the analysis, modeling, and development of robust speech algorithms which address the issues of stress, emotion, and Lombard effect. More details can be found on the CD-ROM, and at the NATO RSG.10 Speech Under Stress web page (http://www.ee.duke.edu/Research/Speech/stress.html).

References


