2007

Speech Under Stress: Analysis, Modeling and Recognition

Sanjay A. Patil, *University of Texas at Dallas*
John HL Hansen, *University of Texas at Dallas*

Speech Under Stress: Analysis, Modeling and Recognition

John H.L. Hansen and Sanjay Patil

Center for Robust Speech Systems, University of Texas at Dallas, Richardson, TX-75080 USA
john.hansen@utdallas.edu

Abstract. In this chapter, we consider a range of issues associated with analysis, modeling, and recognition of speech under stress. We start by defining stress, what could be perceived as stress, and how it affects the speech production system. In the discussion that follows, we explore how individuals differ in their perception of stress, and hence understand the cues associated with perceiving stress. Having considered the domains of stress, areas for speech analysis under stress, we shift to the development of algorithms to estimate, classify or distinguish different stress conditions. We will then conclude with revealing what might be in store for understanding stress, and the development of techniques to overcome the effects of stress for speech recognition and human-computer interactive systems.

Keywords: stress classification, pitch contours, Teager energy operator, robustness in speech recognition, Lombard effect, hidden Markov models, speech technology.

1 Introduction

Speech production involves a sequence of complex coordinated articulator movements, airflow from the respiratory system, and timing of the vocal system physiology. While speech is produced by changes in the articulator positioning, some utterances produced will not be similar in all respects for a speaker. This is because in many situations, the subject is under some type of emotional stress which will impact the utterance causing a deviation in the articulator movements. In human communications, listeners can handle or process these subtle changes far better than the automatic human-machine interface. We have yet to fully comprehend the aspects associated with stress and its effect on human speech production, perception and its impact on automatic speech systems. Thus, speech is a complex signal in a way that encodes information about the speaker, his/her state, acoustic environment, the person's intention, their language background, accent and dialect aspects, and further para-linguistic knowledge.

The main focus of this chapter is to define stress and then move on to show its impact on the speech production system, and thus on the speech systems used

C. Müller (Ed.): Speaker Classification I, LNAI 4343, pp. 108–137, 2007.

for speech recognition and speaker recognition. Specifically when considering robustness in speech recognition systems, the disparity between the training and test utterance significantly impacts performance. An attempt to understand the effect of stress on the human production system will certainly improve the performance of speech recognition systems as well as help in synthesizing speech to simulate emotions.

Before considering analysis and system development, it would be useful to define the "stress" elements of speech. Defining "stress" is a difficult problem because it represents a continuum and is not necessarily a binary decision. In general, a single definition cannot encompass all circumstances. Most definitions might be considered somewhat vague for practical uses. In spite of this, our definition will emphasize aspects from the science of linguistics – *emphasis given to a syllable*. Hence, we use the phrase, "speech under stress" to imply that the subject is speaking under some form of pressure which results in an alteration of the speech production process. The speech occurring under a condition which is devoid of stress, devoid of pressure is termed as "speech under neutral condition". Hence, stress is a psychological state that is a response to a perceived threat or task demand and is normally accompanied by specific emotions (e.g., fear, anger, anxiety, etc). These changes can affect speech behavior, even against an individual's will. Thus, any deviation in speech with respect to the neutral style, whether it is speaking style, word selection, word usage, sentence duration is termed as speech under stress. Therefore, speech under stressful conditions refers to speech spoken under some environmental factor or emotional state which perturbs speech production from a natural, conversational framework. There are many situations where the physical and mental conditions of a speaker can change. Some would include police/fire/ambulance personnel responding to emergency situations, military personnel in either peacekeeping or other military operations. Air traffic controllers represent another group who rely on voice communications in time sensitive stressful conditions. Stress is induced by high cognitive workload, sleep deprivation, frustration over contradictory information, emotion such as fear, pain, psychological tension, and other modern day multi-tasking conditions. Other areas with greater levels of emotions occur include:

1. Forensics – deception detection systems, analysis of 911 phone calls that can include threats [1,2].
2. Safety and Security – air traffic controllers and pilots in noisy high stress environments, deep sea divers, NASA-space explorations, power system operators, [1,3,4,5], military persons facing examination panel [6,7], law enforcement training [8].
3. Psychology – emotional state of patients [9,10].

There have been a number of studies on workload or cognitive task stress and efficiency in noisy environments [3,11].

As shown in Figure 1, speech production and the speaker are affected by various components which include stress caused by cognitive load or physical load, Lombard effect due to the noisy environment, accent change and language

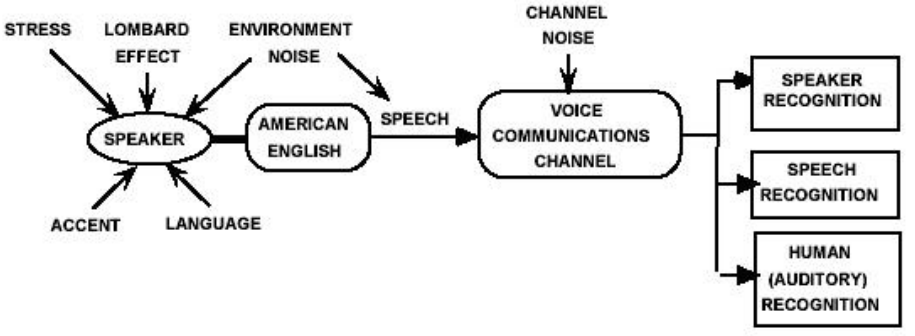

Fig. 1. Effect of stress and other components on speech and speech system

variability. These diverse factors or conditions degrade automatic speech system performance as well as human speech perception. We note that in adverse noisy, stressful situations where speech technology such as speech recognition, speaker verification, or dialog systems are used, addressing noise is not sufficient to overcome performance losses. In noisy stressful scenarios, even if noise could be completely eliminated, the production variability brought on by stress, including Lombard effect has a much more pronounced impact on speech system performance (as will be shown in this chapter).

As voice technology continues to mature, it becomes increasingly important to understand how stress and emotion influence speech production in actual environments. From a communications standpoint, it is clear that there are three distinct domains of speech under stress that include:

(i) physical speech production
(ii) hearing and human perception, including assessing if a subject is under stress, and,
(iii) speech system and technology – feature variation from the acoustic signal for speech and speaker recognition.

2 Domains of Speech Under Stress

2.1 Domain A: Production

Stress is a psychological state that is a response to a perceived threat or task demand and is accompanied by specific emotions (e.g., fear, anxiety, anger). The verbal indicators of stress could be identifying speech markers of stress (e.g., stuttering, repetition, and tongue-slip). Verbal markers of stress range from highly visible to invisible markers as perceived by the listener and that these markers are continuously monitored both consciously and subconsciously by the speaker and thus prone to correction [9].

Respiration is frequently a sensitive indicator in certain emotional situations. When an individual experiences a stressful situation, his respiration rate increases. This presumably will increase subglottal pressure during speech, which

is known to increase fundamental frequency (or pitch) during voiced section [8]. An increased respiration rate also leads to shorter duration of speech between breaths which would affect the temporal pattern (articulation rate). The dryness of the mouth found during situations of excitement, fear, anger, etc., can also effect speech production (e.g., muscle activity of larynx and condition of vocal cords). Muscle activity of the larynx and vibrating vocal cords directly affect the volume velocity through the glottis, which in turn affects fundamental frequency. Other muscles (for example those controlling the tongue, lips, jaw, etc.) shape the resonant cavities of the vocal system and therefore do not have a direct influence on fundamental frequency, though they do contribute to changes in speech production under stress.

It is logical to postulate that if an individual is in a situation where the speed of task completion is critical (e.g., pilot flying an aircraft, air traffic controller), overall duration of utterances would also change under stressed conditions. It has also been suggested that under noisy conditions (Lombard effect), speakers vary their speech characteristics so that portions rich in information are emphasized, and those less important to intelligibility de-emphasized [3,12,13,14,15]. The control of vocal intensity is based on adjustments of laryngeal and subglottal variables. Speakers usually vary their intensity in typical speech production to affect suitable speech intelligibility for human communication [16].

Table 1. Quantifiable / Subjective cues in speech under stress. * The connection between the "observation/feature" and whether it is measurable implies that the stress component can be easily relayed by the speaker and perceived by the listener

OBSERVATION / FEATURE	MEASURABLE
Stuttering, repetition, tongue-slip, pauses between utterances, speed of word production	Quantifiable*
Energy, intensity, pitch (fundamental frequency)	Both quantifiable and subjective
Formant locations / structure (vocal tract), glottal structure (spectral slope), duration	Mostly subjective, but can be measured

Stress has a continuum of observability from the standpoint of the speaker and listener [19]. Changes in speech production which are clearly observable from the speaker, such as a dramatic increase in pitch, are equally observable to the listener. If a speaker wishes to conceal his/her stress, other production changes which are less observable may be altered instead (e.g., on roller coaster rides it is socially acceptable to scream, while in formal speech a speaker may try to maintain pitch and intensity, but adjust less observable markers such as glottal spectral slope). From a communication standpoint, stress could impact the physiological properties of production (i.e., a pilot or person on a roller coaster in high-G force physical movement), environmental factors such as background

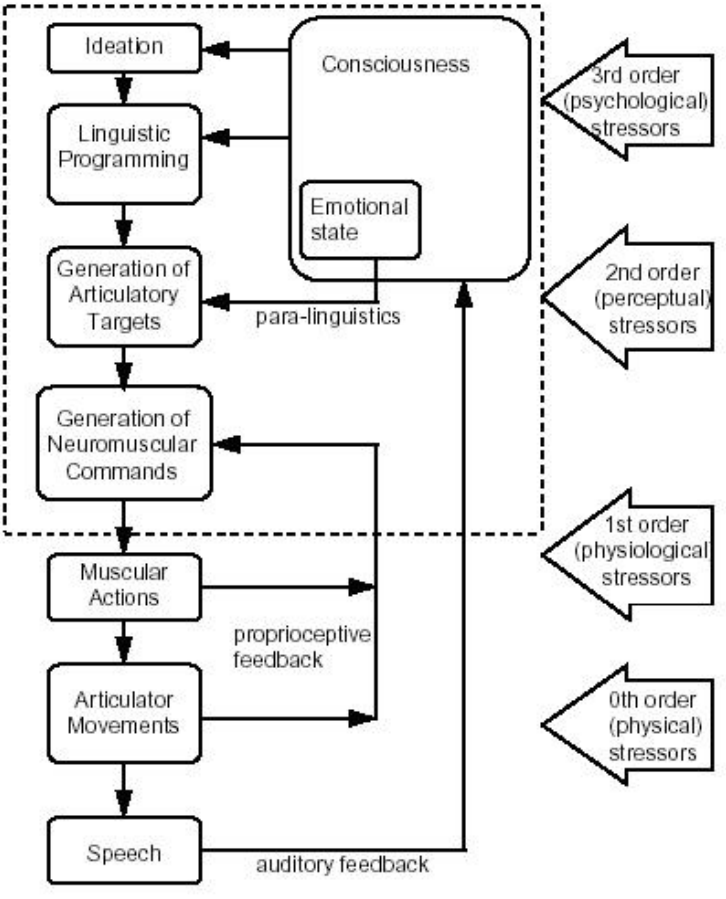

Fig. 2. Effect of stressors on speech production process [17,18]

noise (e.g., Lombard effect), or cognitive factors (e.g., person on the witness stand in a court trial) which could impact word selection.

Therefore, the speech production system can be affected by different stressors which play different roles in the formulation of speech production from word selection, grammar and sentence structure, and physical phoneme/word production. Figure 2 highlights the levels where speech communication/production occurs and their corresponding stress order [17,18]:

1. physical stressors – changes in the vocal apparatus caused due to vibration, movement or G-force, such that it directly affects the articulators.
2. Unconscious physiological stressors – stress causing changes in breathing rate or muscle tension. This might be caused by chemical effects, sleep deprivation or fatigue.

3. Conscious physiological stressors – stress causing increase in vocal effort, increase in voice so as to make oneself heard. This might be due to a noisy environment, experience, or emotion.
4. Internal stress feedback stressors – stress causing changes in vocal effort, mostly caused under the situation which might be interpreted as a threat to one's existence or some perceived conflict/threat.

2.2 Domain B: Perception

Research in speech quality and intelligibility has shown that consonant presence plays a major factor in a listener's ability to perceive the speaker's information content. Therefore, under stressed conditions, a speaker presumably may adjust consonant structure including increasing duration or intensity to give/emphasize additional acoustic cues to the listener. The most important part is that while lexical stress clearly influences duration, the listener will perceive the same utterance with different prosodic content across different stress conditions.

Listeners can identify the "stress markers" in the speaker's message even if these are not obvious. The listener will perceive the signal not merely based on the acoustic signal but using para-linguistics obtained in the context of the conversation, as well as based on his experiences [20]. Therefore, it is important that the speech is perceived in an appropriate manner and that a speaker should insert the appropriate cues within the signal as well as having the listener perceive the utterance in the proper way.

2.3 Domain C: Speech Systems

Speech systems including automatic speech recognition, automatic speaker recognition, speech synthesis, and speech coding / communication systems are all impacted by speech under stress [3,21,22,23]. In the presence of background noise, the speaker will alter his speech in order to communicate more effectively across a noisy environment (this is the Lombard effect). The effect of ambient noise has been suggested to be different for a male speaker versus female speaker [14,15]. In some cases, the situational stress or workload task stress will alter the speech, such as the case if a speaker is experiencing anger, fear, or a pilot flying an aircraft.

As described in Figure 1, the speech signal will be altered by cognitive stress, environmental noise, microphone mismatch which include the speaker, environment, and speech technology employed. These factors all impact the training conditions and therefore will deteriorate the speech systems' performance. If the system is trained with speech from one domain and another one is used for testing, the difference in frequency response causes degradation in the speech system performance. This frequency mismatch can be due to speech production changes caused by stress, microphone mismatch, or communication channel mismatch. Generally speaking, microphone or channel mismatch can reasonably be addressed with a static frequency compensation. Speech under stress however, will require a more intricate compensation scheme over the phoneme sequence.

As part of this chapter, we will focus on speech under stress which also includes the effect of the noisy environment on speech. If we consider a task such as speech recognition, speech under stress will impact robustness. However, for a speech synthesis system, the primary focus is to produce human-like speech, although the ability to impact stress or emotions associated with speech for the synthetic voice can be helpful for some applications. For speech coding, the coding system may not preserve the stress content of the speaker and make this part of communication less effective. For the task of speaker recognition, changes in speaker traits will be difficult to identify and address if the system is trained on neutral data. Hence, a hard binary decision on the type of and extent of stress associated with the speech signal can be helpful in developing more effective speech technology that can be employed in situations where stress / cognitive load / multi-tasking / physical stress / emotion is common place.

3 Analysis

If we consider the range of potential speech characteristics, which could be analyzed for speech under stress, fundamental frequency (or pitch) has historically been the most widely studied. Probably the most extensive early study that focused on analysis was Williams and Stevens [24], while an extensive number of studies have followed since that landmark contribution (see Table 2).

Over the last twenty years, CRSS and earlier variations of our group have performed an extensive level of research on the analysis of speech under stress, algorithm development for detection of stress, speech recognition under stress, and human perception of speech under stress. These studies have concentrated on the SUSAS corpus for the majority of the research [25]. More recently, we have considered other realistic conversational corpora including CU-Move (in-vehicle route navigation dialog), FLETC corpus (police/military training scenario), and UT-SCOPE (speech under cognitive and physical stress conditions) [20,23,26]. The comprehensive feature domains focus on speech production including: fundamental frequency, intensity, duration, formant locations, spectral slope, including an extensive range of features such as traditional MFCCs features and nonlinear TEO-based features.

3.1 Analysis of Fundamental Frequency

Characteristics of fundamental frequency (f_0) include contours, mean, variability, and distribution. A subjective evaluation of more than 400 f_0 contours was conducted across all stress conditions from SUSAS [21,27]. Although f_0 contours indicate excitation differences between styles, they do not reveal significance for particular variations. Moment analysis results, shown in Table 3 include a comparison in mean, variance, standard deviation, average deviation, skewness, and kurtosis. The Student t-test results applied to a pairwise comparison with f_0 data from above show that mean values deviate significantly from neutral as well as most other styles. Speaking styles such as loud and angry showed the

Table 2. Previous studies performed on Speech under Stress

Parameter studied	Analysis
Fundamental frequency contours and its variability	Stress conditions: anger, sorrow, fear [1,2,24]
Mean articulation rate in syllables	Anger, sorrow, fear [1,2,24]
Lombard effect speech	[14,15]
Vibration space shift rate (VSSR) from speech spectrograms	Fundamental frequency [3]
Monitoring heart rate and spectral centroid of first formant	Need for further research [20]
Pitch, amplitude, timing measurements	Elevated pitch and amplitude, and increased variation [5]

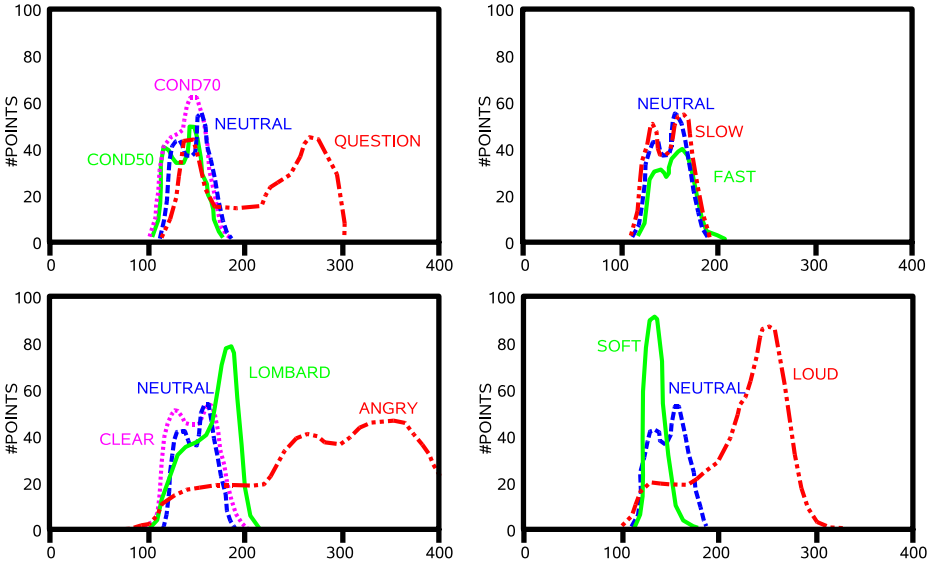

Fig. 3. Fundamental frequency (pitch) distributions across different speaking styles and stress conditions

widest deviation from neutral. Mean fundamental frequency is a good indicator over a wide variety of stress conditions. Loud, angry, and Lombard mean fundamental frequency are all significantly different from neutral as well as all other styles considered.

From Table 3 and Figure 3 and F-test statistical analysis, we can concur for most cases, f_o variance is shown to be significantly different from neutral as well as many other styles, and therefore is a good differentiating stress parameter. Pitch variance is not reliable for moderate versus high computer workload task (COND50 vs. COND70) conditions, and for slow and fast stress speaking conditions. Though the pitch distributions from Figure 3 are generally bimodal,

Table 3. Analysis of Fundamental frequency over various speaking styles and stress conditions

Stress Condition	Mean Value	Max. Value	Min. Value	Ave. Dev.	Stand. Dev.	Var.	Skew.	Kurt.
Neutral	142	182	116	13.4	15.4	239	0.22	-0.98
Slow	140	174	114	12.7	14.6	212	0.27	-1.10
Fast	149	186	121	11.9	14.0	195	0.19	-0.80
Soft	135	267	114	5.2	9.7	93	7.02	86.5
Loud	209	276	113	37.9	44.1	1944	-0.54	-0.97
Anger	283	400	96	44.3	56.3	3166	-0.38	0.44
Clear	150	211	103	19.1	22.1	489	0.23	-0.94
Cond50	140	205	111	13.7	16.2	263	0.41	-0.25
Cond70	143	216	111	13.7	16.3	266	0.34	0.01
Lombard	163	229	109	21.6	24.8	614	-0.25	-1.05

in certain stress styles (angry, question, soft, loud) the shape does deviate significantly from neutral (as measured by Kolmogorov-Smirnov pairwise test for distribution). We should note that contour shape of course plays a major role for question style. We therefore conclude that while a range of f_0 factors change in stress speaking styles, mean and variance can be effective traits for stress classification.

3.2 Analysis of Duration

In a previous study, it is shown that for stress conditions where time is of the essence, word duration, as well as subword durations such as changes in vowels versus consonants, and consonant presence, plays a major factor in a listeners' ability to perceive the speaker's information content [3,21].

As seen in Table 4, mean word duration as expected, increases for slow and decreases for fast spoken speech. The duration of consonants, semivowels, and diphthongs (to a lesser degree) remain constant in soft versus loud conditions, vowel duration decreases slightly for soft speech, and increases significantly for loud speech (as well as angry speech).

Upon more fine analysis, we observe significant changes in mean word duration for several stress conditions and proposed that it could be possible that overall word duration remains constant with shifts between consonant and vowel sections. Mean vowel and consonant duration possess similar discriminating abilities. Similarly, for variance, vowel and consonant classes continue to have reliable stress discriminating power.

Since vowels and consonants show major changes across all stress styles, several proposed discriminating features were proposed. The derived features are consonant versus vowel duration ratio (CVDR), consonant versus semivowel duration ratio (CSVDR), and vowel versus semivowel duration ratio (VSVDR) for all the stress styles. Table 5 summarizes the results which illustrates shifts in overall word duration, as well as movement between vowel, semi-vowel, and consonant classes [3,21].

Table 4. Word and Speech Class Duration over various speaking styles and stress conditions

Stress Condition	Mean Duration (msec)									
	N	Sl	F	So	L	A	C	C50	C70	Lom
Word	478	827	353	509	650	662	666	482	501	572
Vowel	160	294	115	147	253	271	202	148	147	198
Consonant	71	107	52	87	73	62	128	79	86	73
Semivowel	60	126	57	71	76	85	83	71	68	97
Diphthong	192	374	147	210	294	315	199	176	178	249
Stress Condition	Duration Variance (msec)									
	N	Sl	F	So	L	A	C	C50	C70	Lom
Word	18	49	12	16	28	41	40	16	14	24.0
Vowel	7.9	21	3.6	6.2	19	23	17	7.6	7.6	13.0
Consonant	1.8	7.1	1.1	2.9	3.7	3.3	10	2.5	3.3	2.6
Semivowel	0.7	7.1	1.0	1.3	2.9	7.8	3.2	1.7	1.4	4.3
Diphthong	3.3	14	1.0	1.1	5.6	7.0	3.3	2.4	3.4	3.5

Stress Style Key:
N – Neutral, Sl – Slow, F – Fast, So – Soft, L – Loud, A – Angry, C – Clear, C50 – computer task Cond50, C70 – computer task Cond70, Lom – Lombard effect

CVDR and CSVDR suggest that there is a shift in the percentage of time spent in vowels and semivowels towards consonants for soft, clear, and to a lesser degree the two computer task conditions (COND50 and COND70). These results indicate that the presence of stress influences word and individual phoneme duration characteristics.

3.3 Analysis of Intensity

Next, we consider analysis of intensity over stress speaking styles at the word level and phoneme levels. To focus the intensity analysis on the core portion of each phoneme, the phoneme boundary was reduced by 10% from both directions towards the phoneme mid-point, with RMS energy found for each phoneme and overall word. As expected, a marked increase resulted for loud and angry conditions, while soft, clear, and speech under the two computer task workloads had reduced word intensity. Vowel intensity remained constant for slow, clear and Lombard conditions, while consonant intensity increased for soft, angry, question, and speech under two computer task workloads (see Table 6). Word intensity possessed a good level of stress discriminating ability. However, experiments show that for several stress styles, such as angry, duration and intensity are interrelated. For detection of stress, mean RMS intensity is as successful

Table 5. Analysis of Duration over various speaking styles and stress conditions

Stress Condition	Analysis of Mean Duration (msec) and Ratios						
	Word	Vowel	Semivowel	Consonant	CVDR	CSVDR	VSVDR
Neutral	478	166	59.6	70.6	0.426	1.184	2.777
Slow	827	308	126	107	0.349	0.850	2.437
Fast	353	964	57.4	51.8	0.429	0.901	2.100
Soft	508	158	70.9	87.3	0.552	1.231	2.230
Loud	650	260	75.5	72.6	0.279	0.962	3.444
Anger	662	279	84.6	62.1	0.223	0.734	3.294
Clear	666	201	82.9	128	0.634	1.539	2.429
Cond-50	482	153	71.4	78.8	0.516	1.103	2.136
Cond-70	501	152	67.9	86.0	0.566	1.267	2.239
Lombard	572	207	97.3	73.1	0.353	0.750	2.214

as mean duration. Intensity variance across words or phoneme classes were not consistently successful for stress detection.

3.4 Glottal Pulse Shaping

The spectral based characteristics from the glottal source and vocal tract response are also impacted during speech production under stress. In this subsection, we focus on glottal source changes and in the subsequent section on vocal tract response characteristics. Glottal spectral source factors which include spectral slope, center of mass, and mean spectral level were analyzed as potential acoustic correlates of speech under stress [21].

For glottal flow spectra under all the ten stress conditions, the shape of glottal flow spectra are similar but differentiating features include spectral slope and amplitude [12]. Typically, for Lombard and angry styles, variability in spectral amplitude was observed in the 2 to 4KHz band. This generally implies a change in the shape of the glottal pulses under these conditions.

Using linear regression, the spectral tilt information was extracted across the glottal spectrum. Figure 4 summarizes that all speaking styles have a spectral tilt significantly different from neutral. Based on spectral characteristics, under certain stress conditions (loud, angry, and Lombard), glottal pulses will have steeper slopes with sharper glottal pulse corners (or irregular shapes) caused by a combination of changes in sub-glottal air pressure, vocal-fold tension, and uneven or sudden closure of the vocal folds during phonation. Alternatively, for slow and soft styles, glottal pulse shape will have gradual rise and fall times and overall smooth shapes, resulting in reduced energy in high frequency content and a steep spectral slope (-15dB/octave).

The analysis of glottal source spectrum revealed that parameters such as spectral slope and the distribution of energy to be important for relaying stress.

Table 6. Mean and Variance for Word and Speech Class Intensity over different speaking styles and stress conditions

Stress Condition	Mean Intensity (RMS)									
	N	Sl	F	So	L	A	C	C50	C70	Lom
Word	7663	7982	7812	7277	10561	11307	7067	7075	6934	8286
Vowel	9610	9692	9404	9326	12002	12700	9786	8857	8996	9699
Consonant	1394	1481	1425	1866	1164	1562	1287	1592	1715	1401
Semivowel	10032	9323	9983	10072	9443	11629	8272	8498	8353	8322
Diphthong	10125	9989	10460	9393	14800	14724	10394	9807	9742	10913
Stress Condition	Variance of Intensity (RMS)									
	N	Sl	F	So	L	A	C	C50	C70	Lom
Word	12.3	8.5	10.1	16.0	31.2	50.7	6.5	9.2	8.3	16.8
Vowel	93.3	76.5	93.4	63.6	116	193	109	92.2	101	80.4
Consonant	21.8	32.1	20.1	35.8	26.0	33.6	24.9	24.1	33.1	23.9
Semivowel	128	106	136	102	231	571	75.7	162	187	152
Diphthong	38.7	14.6	29.0	22.1	17.3	19.5	23.6	23.2	30.8	8.4

3.5 Vocal Tract Spectrum

To study the effect of speech under stress on the vocal tract spectrum, the mean, variance, and distribution of formant location and bandwidth across extracted phonemes were analyzed [21].

The previous research found shifts in frequency content for subjects performing a timed arithmetic task [12,28]. The results were more pronounced for front vowels than back vowels, with weaker third and fourth formants (reduced amplitudes) for the stress versus control conditions. We have found that when a speaker is under stress, typical vocal tract movement is effected, suggesting a quantifiable perturbation in articulator position.

Slow, loud, angry, and clear speaking styles show the widest shift in F1 formant locations. F2 formant frequencies generally increase among most conditions. Only slight changes occur for F3 and F4 locations across all styles. Formant bandwidth show large variations in mean for the first two formant frequencies with some changes for F3 and F4. The variance of formant location and bandwidth also showed shifts, with especially large changes in variance for loud, angry, and clear styles for F1. The changes in formant structure (location and bandwidth) are seen in Figure 5 and Figure 6 for Lombard, angry and neutral styles.

A series of Student T-tests were performed assuming both equal and unequal variance. Mean shifts in formant location (F1, F2, and F3) for loud, angry, and clear were significantly different from neutral. It was seen that styles which vary in formant location, will also increase formant variability in conveying that stress condition.

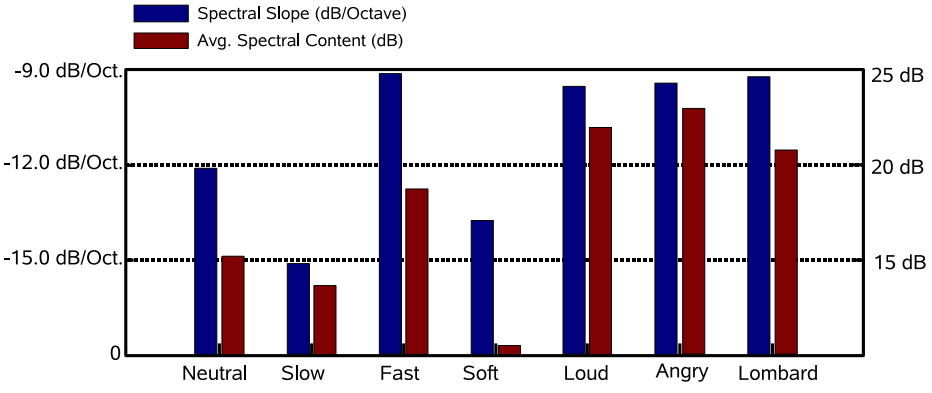

Fig. 4. Special Tilt (Glottal Pulse Shaping) Left Axis: spectral slope in dB/Octave, Right Axis: average spectral content in dB

Average formant location, average formant bandwidth and the variance of these all display varying degree of stress relayer information.

4 Applications

As discussed in the previous sections, many environmental and situational factors contribute to variation in speech production. Studies have shown that speech produced under stress causes significant loss in performance for traditional speech recognition algorithms. Stress and emotional characteristics must also be captured and modeled in order to produce more natural sounding speech coding and text-to-speech synthesis techniques. The importance in understanding how speakers vary their production systems to convey emotional or task induced stress has been shown in the previous section.

In the past, limited research has been conducted on the effect of stress on speech systems. Based on our investigations, the speech aspect that appears to provide the clearest indication of emotion or stress is fundamental frequency over time. Although important for the analysis of speech under stress, variation in pitch may not be a critical factor in attempting to reduce errors in traditional speech recognition algorithms. However, if the analysis of such parameters were to show statistically reliable indicators, it may be possible to formulate front end analysis procedures to identify periods of high stress. Recent studies demonstrate the potential for reliable stress classification via nonlinear, articulatory, and speech production features [9,24,29,30,31]. Once a period of speech under stress has been identified, a recognition system incorporating a compensation procedure specific to that form of stress could be used [32,33,34,35].

Although some variation in duration may not seriously affect speech systems, if the phonemes used for discrimination decreases in length, the probability of word misclassification can increase. Similar problems could arise for an increase

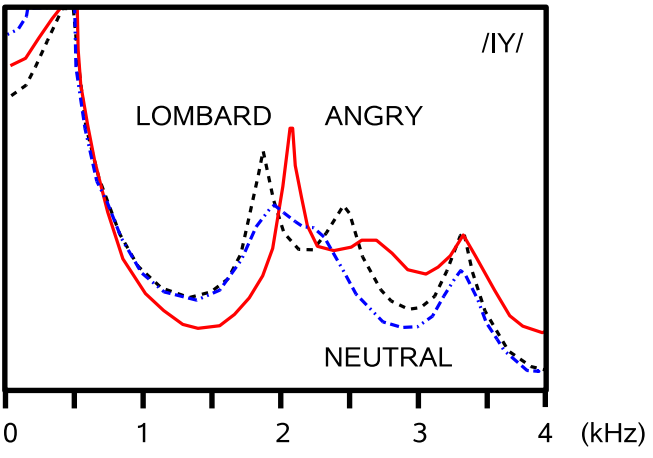

Fig. 5. Vocal tract spectrum for /IY/ phoneme

in duration in HMM modeling due to the finite number of states and numerical accuracy available in computing state transition probabilities.

Today, commercial based speech recognition systems can achieve more than 95% recognition rates for large vocabularies in restricted paradigms with relatively noise-free environments.

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. The main reason for this is that the performance of existing recognition systems which assume a noise-free tranquil environment (or train-test matched conditions), degrade rapidly in the presence of noise, distortion, and stress.

It is suggested that algorithms that are capable of detecting and classifying stress could be beneficial in improving automatic recognition system performance under stressful conditions. Furthermore, there are other applications for stress detection and classification. For example, a stress detector could be used to detect the physical and/or mental state of a pilot and that detection could put special procedures in place such as rerouting of communications, redirection of action, or the initiation of an emergency plan. To be able to detect and classify stress, it is necessary to understand the effect of stress on acoustical features.

There are two processing stages in a stress detection system. In the first stage, acoustical features are extracted from an input speech waveform. The second stage is focused on detection of stressed speech from neutral using one or more available methods.

A variety of methods exist for stress detection which include, but not limited to, detection-theory based methods, methods based on distance measures, and statistical modeling based techniques. A representative sample are presented in this section. These methods include:

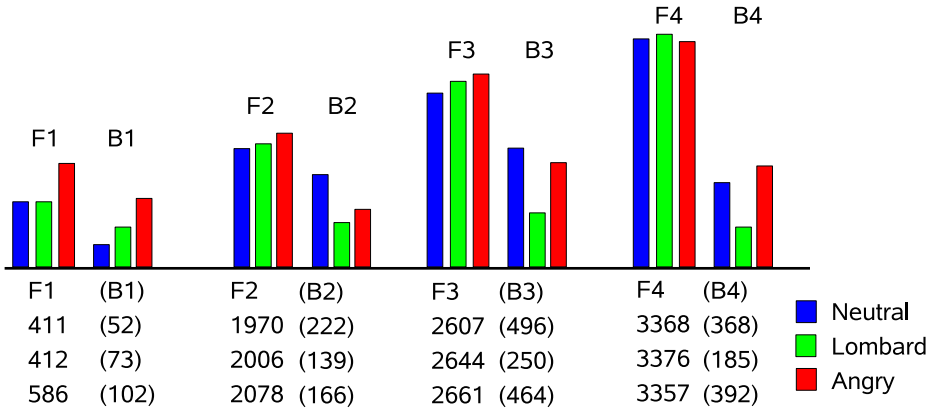

Fig. 6. Formant Frequency Location and Bandwidth value and distribution

(i) Neural Networks with linear speech model-features,
(ii) Optimum Bayesian detection used for stress classification,
(iii) TEO-based nonlinear speech features for both stress classification and stress assessment.

In the next section, we first focus on speech recognition in section 4.1 followed by stress detection methods in section 4.2.

4.1 Speech Recognition

To improve the performance of speech recognition systems in stress and noise, a number of methods have been considered including multi-style training, simulated stress token generation, training and testing in the same noise. While these methods help in matched conditions, the results degrade as test conditions drift from the base train condition. Some methods which address this drawback focus on estimation of speech features in noise, adapting speech enhancement techniques, and / or incorporating stress equalization [13,32,36]. The concept of stress equalization is based on a processing scheme which operates on a parameter sequence that is extracted from the input speech under stress. The stress equalization algorithm attempts to normalize the variation of the parameter sequence due to the presence of stress on the input speech signal.

Stress equalization techniques are a front-end processing approach to improve speech recognition under stress. The techniques can rely on maximum likelihood compensation factors to project the input stress modified features into a neutral-like space, where a neutral trained automatic speech recognition system is used. Figure 7 illustrates the impact of stress and noise on speech recognition performance [3,25]. We see that a basic speech recognition task in neutral, noise-free conditions is significantly impacted by the presence of stress (e.g., an average 31% reduction in recognition accuracy), and stress and noise combined (e.g., an

average 58% reduction in recognition accuracy). Lombard, loud and angry stress styles in noise are significantly impacted. To address stress and noise, a combination two-tier approach was considered based on maximum likelihood stress compensation algorithm directly on the speech features, combined with noise suppression using Auto-LSP constrained iterative speech enhancement [3,37]. This combination scheme provided measurable levels of speech recognition improvement over noisy stressful conditions (see Figure 9, average accuracy improvement of 27%). A more rigorous stress compensation scheme developed for MFCC cepstral parameters was shown to have an even greater performance improvement in noisy Lombard effect speech [33].

Due to the extensive level of research activity in robustness for automatic speech recognition in stress and noise, it is not possible to consider even most of the advances over the past 15 years. The overview study [25] provides an effective and comprehensive step, and the interested reader is encouraged to consider the extensive bibliography at the end of this chapter. Our intension here is to provide a brief overview of the research topic in automatic speech recognition.

Another way to compensate for stress is to use a front-end artificial neural network. Figure 8 illustrates the use of an artificial neural network (ANN) for improving the performance under noisy stressful speech conditions. With a feature enhancement ANN (FE-ANN), a unique FE-ANN is created for each keyword model and further evaluated using a semi-continuous HMM recognizer followed by a likelihood ratio test for keyword detection [12,25,38]. The results show that a front-end ANN can provide consistent improvement for keyword recognition under Lombard effect.

A more rigorous method to address stress was based on morphological constrained feature enhancement with an adaptive cepstral stress compensation technique would be a third alternative studied for speech recognition systems [33]. Figure 10 shows the improvement achieved with MCE with adaptive mel-Cepstral Compensation. It should be noted that some features which are robust for speech recognition in noise, may not be as successful in stress and those successful in stress may not be as successful in noise. For example, linear predictive (LP) based MFCC features are more effective for speech recognition under stress versus FFT based MFCC, FFT based MFCCs are more successful for speech recognition in noise but performance decreases for speech under high stress conditions (angry, loud, etc.) [36]. The studies here suggested that effective speech features, compensation methods, and alternative training methods, can all lead to improved speech recognition performance in speech under stress.

4.2 Stress Detection

Since the range of speech under stress can include several broad types, the domains for stress detection in partitioned into the following four categories:

1. Speech under deception
2. Lombard effect detection

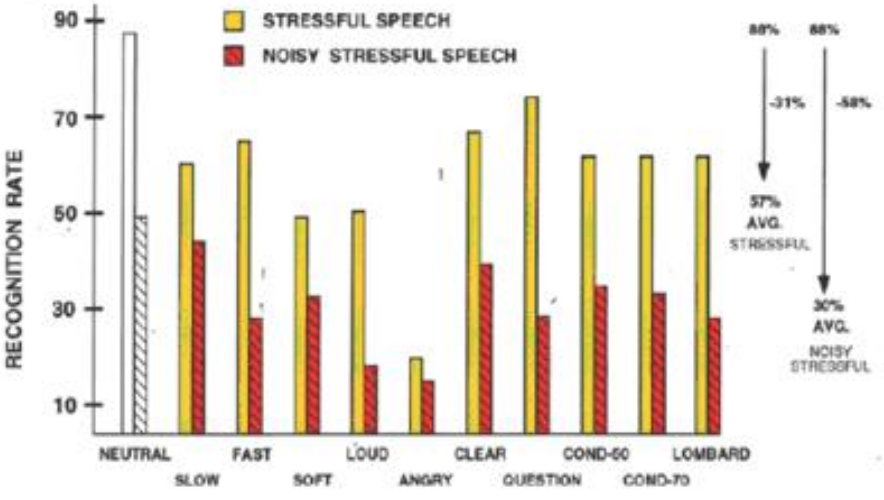

Fig. 7. Application of stress equalization for ASR - VQ-HMM ASR system

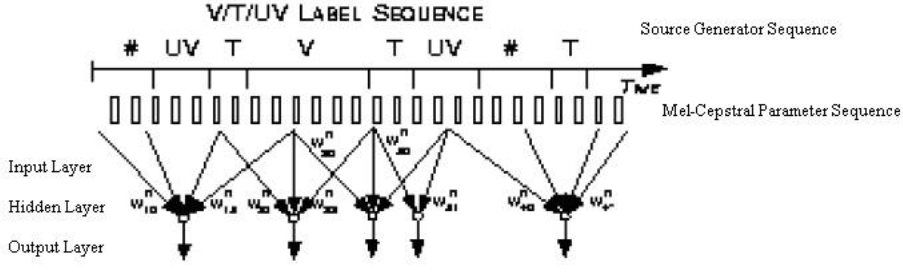

Fig. 8. FE-ANN for robustness of Speech Recognition Systems

3. Cognitive Stress detection
4. Physical Stress detection

Certain systems use voice stress analyzers based on microtremors, which have been shown to not be good indicators of stress [39,40].

4.3 Detection-Theory-Based Framework for Stress Classification

A Flexible framework for stress detection can be easily established using detection theory. For such a scheme, there are two hypotheses termed H_0 and H_1. Under H_0, the speech is neutral; while under H_1, the speech is stressed. Given an input speech feature vector, x, two conditional probability density functions (PDF), $p(x|H_0)$ and $p(x|H_1)$, must be estimated. With these PDFs, the likelihood ratio, λ, is defined as follows,

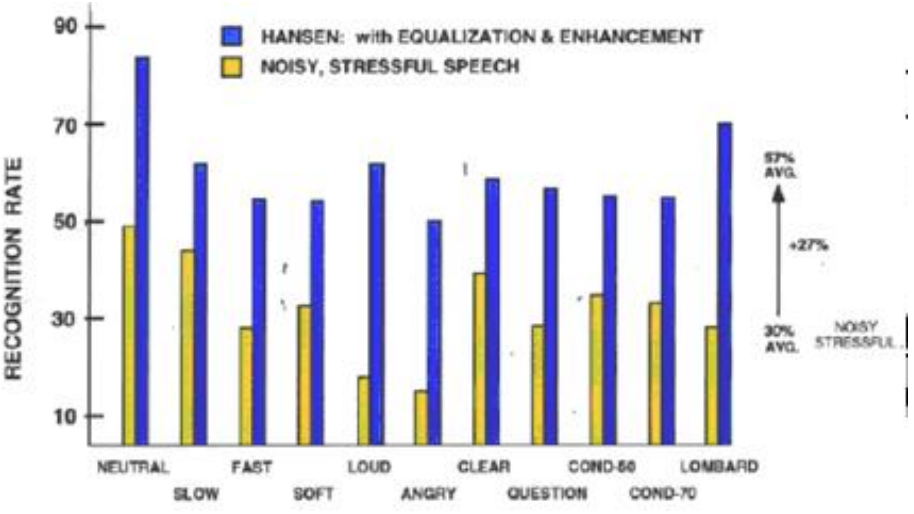

Fig. 9. Robust Recognition under noisy and stressful speech [3,27]

$$\lambda = \frac{p(x|H_1)}{p(x|H_0)} \tag{1}$$

The decision of whether the input speech is neutral or stressed is made by comparing the likelihood or log likelihood ratio with a predefined threshold, β. If the ratio is larger than β, the input speech is detected as stressed; otherwise the input speech is classified as neutral. The value of β depends on the particular criterion used for detection.

4.4 A Distance Measure for Stress Classification

The detection of stress versus neutral speech can also be achieved using a distance measure. For a given input observation speech feature vector and two prior feature distributions (one for neutral, and one for stress), two distance measurements can be obtained: the distance between the given vector and neutral speech distribution, along with the distance to the stressed speech distribution adjusted for variance. This distance measure reflects the proximity of the input sequence to the distribution of general neutral or stressed speech feature data.

Previous CRSS studies have concluded that using individual speech features for stress detection show a range in detection performance as summarized in Table 7. Acoustical features such as duration, intensity, pitch, glottal source information, and formant locations for vowels were studied for stress detection performance using isolated words from the SUSAS corpus. The two methods for detection include a traditional binary hypothesis detection-theory method, and a dual PDF distance based method. Table 8 shows the results for stress detection performance as the number of feature observations for detection is increased

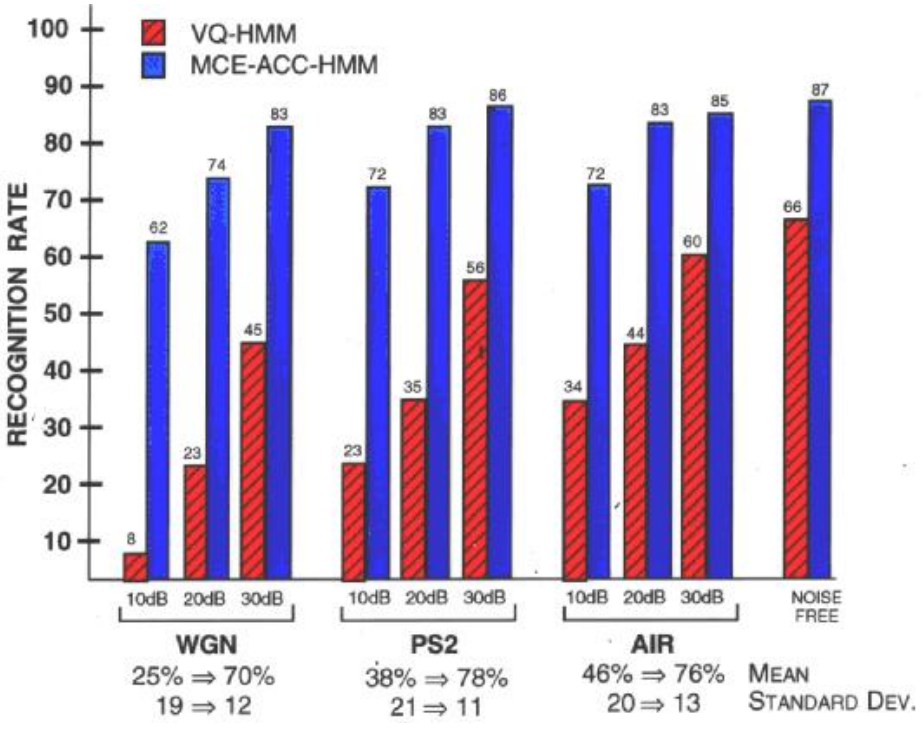

Fig. 10. MCE-HMM based Robust Speech Recognition system for stressful speech under noisy conditions

from 1 to 10 [41]. In general, given the error rate levels for the three stress classes tested, extensive experimental evaluation of stress detection at CRSS, we conclude the following:

1. Vowel duration is not a good feature for stress detection.
2. For intensity, increasing the input vector length does improve performance, especially for detecting angry and loud speech based on a detection-theory algorithm. As for the distance measure approach, increasing the input vector length does not always improve performance. The open-set test results also show that both methods perform better for detecting angry and loud speech versus detection for Lombard effect speech.
3. Compared to duration and intensity, pitch has much better performance for stress detection. Either of the methods perform similar with pitch feature.
4. The open-set results from the detection-theory-based method show that spectral slope (indicator of glottal source) is more suitable for detecting angry speech than for detecting loud speech with Lombard effect from neutral speech.
5. The features representing the vocal tract spectrum – formant location, are not suitable for stress detection.

Table 7. Stress Detection Studies using Traditional Features (Stress Conditions: Lombard, loud, angry)

Feature set	Stress/ Neutral Error Rates
Pitch	6-21 % variation
Glottal Spectral Slop	18-36 %
Intensity	18-36 %
Phone Duration	28-46 %
Formant Location	
1st Formant	38-46 %
2nd Formant	50-58 %
Feature Fusion	
Duration + Intensity + Mean Pitch	0-17 %

Table 8. Error Rates (%) of Open-set Pairwise Stress Classification using the combination of mean pitch, duration, and intensity as the feature

Vector Length	Speaking Style of Submitted Test Speech						Overall Error Rates	
	Neutral	Angry	Neural	Loud	Neutral	Lom.	Mean	Std. Dev.
1	17.68	17.03	11.67	11.97	19.85	21.21	16.5 %	3.97
5	6.15	5.00	4.62	4.62	13.08	13.08	7.76 %	4.16
10	1.67	0.00	3.03	3.03	13.64	16.67	6.34 %	6.98

The results from this section provide a representative perspective on the use of traditional speech production features for stress detection. Further studies have focused on the fusion of multiple features, and the interested reader is encouraged to explore the following references [6,7,8,9,17,30,31,41,42,43].

4.5 Neural Network Based Systems

Neural Network classifiers can also be employed for stress classification. A neural network based classification algorithm was considered for stress classification using cepstral-based features which have traditionally been employed for recognition [30]. Mel-cepstral parameters represent the spectral variations of the acoustic signal. It is suggested that such parameters are useful for stress classification since vocal tract and spectral structure vary due to stress.

Frame-based and word-level features performed in the ranged from 11-17 % for a 35 word test set which is greater than chance (9 % for eleven stress types in the SUSAS corpus). Most importantly, some stress conditions had reasonably good classification performance [30].

Another study considered the most effective feature subset for each targeted stress condition determined during a training phase emphasizing the most discriminating features (out of 27 studied) for classification of each stress style [30]. It has also been shown that a multi-dimensional HMM based system can

be formulated which combined stress classification along with automatic speech recognition [44]. The resulting N-Dimensional HMM system resulted in a 73.8 % reduction in error rate as compared to the single channel stress dependent isolated word recognition system.

4.6 Stress Classification Using Nonlinear Speech Features

Next, stress classification can be considered from an alternative speech production modeling perspective. The assumption that airflow propagates as a plane wave in the vocal tract may not be the most accurate airflow model of speech production, since the flow is actually separated with concomitant vortices that are distributed throughout the vocal tract. Teager pioneered alternative approaches to speech modeling and also suggested that hearing could be viewed as the process of detecting the energy [45,46,47,48,49]. Over the past ten years, a number of studies have suggested that the so called Teager Energy Operator (TEO) can be employed to formulate new features for stress classification [6,7,8,41,42,43,50].

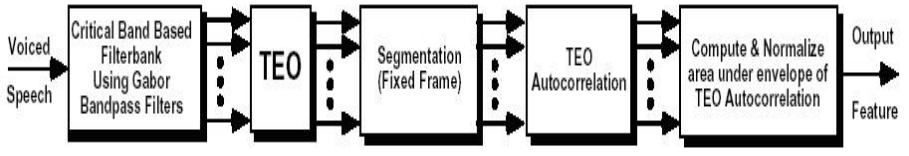

Fig. 11. Flow Diagram of TEO-CB-AutoEnv based feature

One of the effective nonlinear TEO-based feature developed by Zhou, Hansen and Kaiser [41,42,43,51] is the TEO operator, partitioned across a critical frequency band with an autocorrelation envelope analysis performed. The flow diagram for the TEO-CB-AutoEnv is shown in Figure 11. The theory is that the autocorrelation envelope is able to track the variability / regularity of the fine energy structure reflected in the TEO critical band partition, a trait which occurs in speech production for high stress conditions. Results from an evaluation using speech material from the SUSAS corpus is shown in Figure 12. Stress classification performance is significantly better than traditional MFCC based spectral features, or excitation based f_0 (pitch) information [41]. The performance is consistent for neutral versus emotion, speaking style, Lombard effect, and actual roller coaster ride speech under stress. The feature therefore is effective in both simulated and actual stress speech scenarios [41,42,43,51].

The same TEO-CB-AutoEnv feature has also been employed for stress detection in other scenarios. Figure 13 shows results for stress detection using data from a military examination task (SOM – Soldier of the Month Training). The results show as significant reduction over an MFCC feature based HMM baseline classifier using the new TEO-CB-AutoEnv feature [6] based on single word "no" decisions. Further experiments on this same SOM corpus have explored the impact of increased test token duration. The results from Figure 14 show

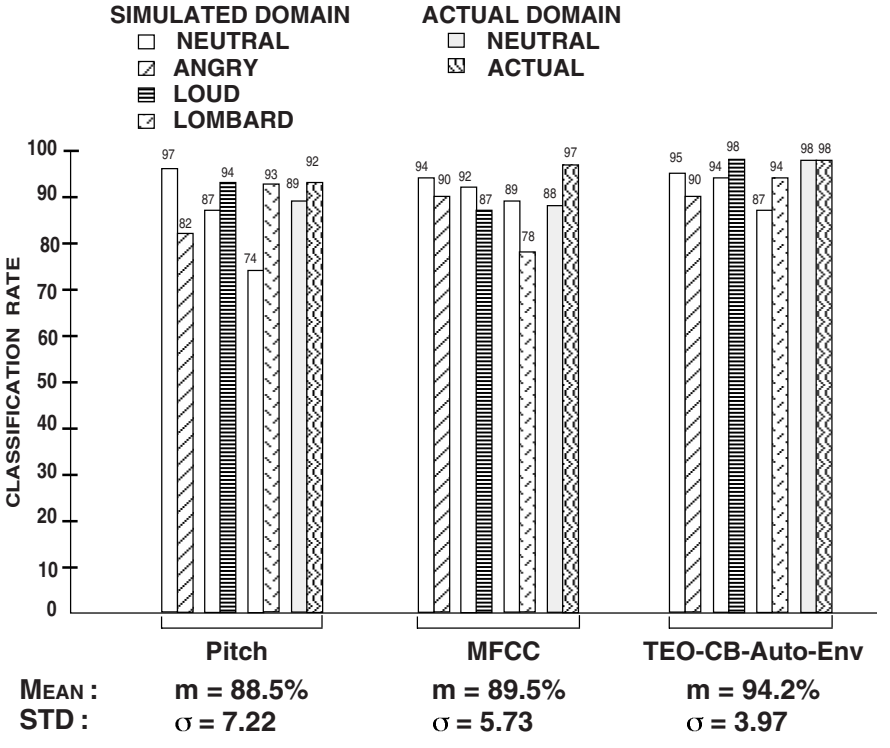

Fig. 12. Results comparing performance of TEO-based system with traditional features

that if 0-40 % of the vowel duration is removed, stress detection performance is maintained.

More recent studies employing real-life speech corpora such as US Army SOM (Soldier of the Month) or FLETC (law enforcement training scenario involving hostage rescue with weapons) have shown that TEO based features can be used for stress detection, as well as stress assessment over defined time periods [6,7,8,52,53].

4.7 Synthesis and Conversion of Speech Under Stress

As seen in our studies and elsewhere as well, the measured features that can reflect stress include changes in pitch and other excitation features, word/phoneme duration intensity, and spectral content. To activate the desired stress intonation for synthesized speech requires that the necessary variations in the actual stressed speech be represented in voice quality, pitch and duration of individual phonemes within the utterance [17,32,36,54,55,56,57,58,59]. This helps improve the naturalness of the synthetic speech. Previous approaches directed at integrating emotion in text-to-speech synthesis systems have concentrated on formulating a set of fixed rules to represent each emotion [58,59]. To represent a

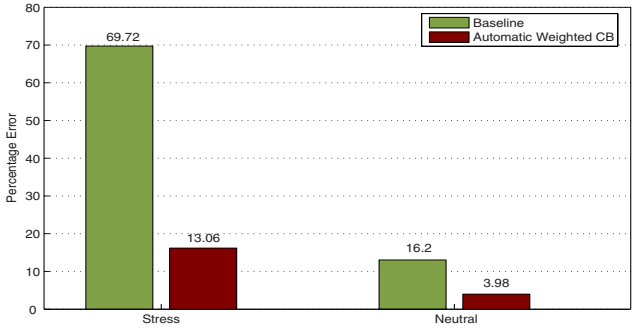

Fig. 13. Classification Error rates for Open Speaker Set (based on a single word "no" for SOM database)

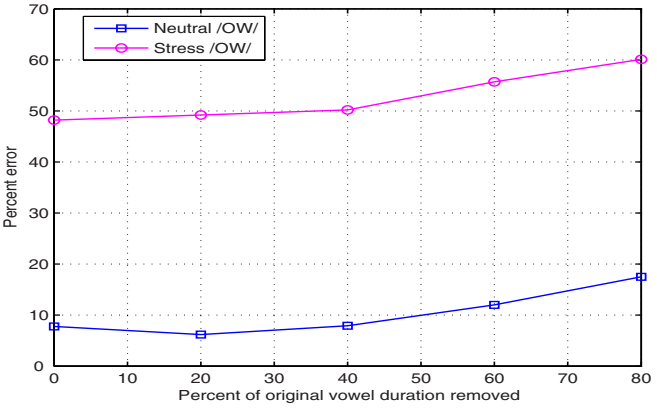

Fig. 14. Effect of vowel duration on stress classification error rates

range of variations for continuous speech, a fixed set of rules is not sufficient in general if we wish to have natural sounding speech.

It is possible to impart stress onto existing speech. One proposed method focused on converting CELP based excitation and vocal tract spectral structure from neutral to produce stress speech (Lombard, loud, angry speech styles) [32]. A subsequent study focused on HMM based modeling for voice conversion of neutral to stressed speech. The model showed it was possible to model stress perturbation techniques from one set of speakers and successfully impart these changes onto new neutral speakers [55].

Further studies have also explored the ability to impart emotion onto synthetic speech for text-to-speech applications [1,5,6,26,60,61,62,63]. These methods can be viewed as imparting a caricature or exaggerated version of the emotion/stress in order to make the emotion obvious to the listeners, and therefore generally do not always reflect true speech production changes that are more subtle. Further

research is necessary to better understand speech under stress for synthesis, as well as perception of stress for synthetic speech applications.

4.8 Speech Coding System

As for speech coding, preserving the naturalness of the speech on the receiver side would help convey the emotional or stress state of the speaker. The stress perturbation algorithm for CELP coding system modified the pitch, gain, and the formant locations to convey the emotional state of the speaker [32]. The method along with hidden Markov model demonstrated conservation of speaking styles for isolated words under neutral, loud, angry and Lombard effect speaking conditions [32,55]. Future development of speech coding algorithms need to effectively capture the changes in speech production under stress. New advances in alternative excitation modeling based on GEMS or p-mike could offer improved techniques to encode speaker stress state for voice coding applications.

5 Discussions and Future Directions

As speech and language technology continues to mature, the need to effectively analyze, model, encode, detect, and classify speech under stress will increase significantly. Voice interactive systems including dialog and human-machine systems can benefit from knowledge of the speaker state. This information can help improve technology for speaker and speech recognition providing systems that are more effective in actual multi-task scenarios. The challenge, however, is to employ a framework which can provide effective analysis and modeling for improving such speech technology.

The source generator framework (SGF) proposed in [27,29,33] offers an effective means of modeling deviations from neutral to stress, and has been employed for a variety of stress equalization methods [17,25,29,33]. The basic structure is represented as:

$$(\text{speech})_{stress[X degree]}(\text{feature set}) = \Psi[(\text{speech})_{neutral}(\text{feature set})] \quad (2)$$

where Ψ [] is the transfer operator function which transforms neutral to stressed speech which has a certain degree of stress, say X. The above problem is two fold,

1. To define (in quantifiable sense) the degree of stress, X.
2. To define the speech production transfer operator Ψ[].

We model the transformation Ψ[] of the speech features in the neutral domain to an output stress domain. Prior formulation considered Ψ[] operators in the pitch, duration, intensity, glottal source, and vocal tract spectrum domains. It is important to recognize that if the transfer operator function is dependent only on the stress and phoneme, and generally independent of the speakers, it can be applied in more scenarios. An inverse transformation is therefore developed

using this structure to compensate for the presence of stress. It is suggested that future advances in stressed speech processing could be realized using the Source Generator Framework, resulting in more effective speech and language technology with sustained performance in adverse speech/noise/environmental conditions.

References

1. Alm, C.O., Roth, D., Sproat, R.: Emotions from Text: Machine Learning for Textbased Emotion Prediction. In: Proceedings of HLT/EMNLP 05, Vancouver (2005)
2. Hollien, H.: Forensic Voice Identification. Academic Press, London (2002)
3. Hansen, J.H.L.: Analysis and Compensation of Stressed and Noisy Speech with Application to Robust Automatic Recognition. PhD thesis, School of Electrical Engineering, Georgia Institute of Technology, Atlanta (1988)
4. Simpson, C.A.: Speech Variability Effects on Recognition Accuracy Associated With Concurrent Task Performance by Pilots. Technical report, Psycho-Linguistic Research Associates (1985)
5. Sproat, R., Olive, J.: Text-to-Speech Synthesis. In: Rabiner, L., Cox, R. (eds.) IEEE/CRC Press Handbook of Signal Processing, CRC Press, Cleveland (1997)
6. Prahallad, K., Black, A., Mosur, R.: Sub-Phonetic Modeling for Capturing Pronunciation Variation in Conversational Speech Synthesis. In: Proceedings of the 31th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06), Toulouse (2006)
7. Ruzanski, E., Hansen, J.H.L., Meyerhoff, J., Saviolakis, G., Koenig, M.: Effect of phoneme characteristics on TEO Feature-based Automatic Stress Detection in Speech. In: Proceedings of the 30th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05), Philadelphia, vol. 1, pp. 357–360 (2005)
8. Rajasekaran, P.K., Doddington, G.R., Picone, J.W.: Recognition of Speech under Stress and in Noise. In: Proceedings of the 11th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '86), Tokyo, pp. 733–736 (1986)
9. Cairns, D.A., Hansen, J.H.L.: Nonlinear Analysis and Detection of Speech under Stressed Conditions. Journal of the Acoustic Society of America 96(6), 3392–3400 (1994)
10. Dharanipragada, S., Rao, B.D.: MVDR-based Feature Extraction for Robust Speech Recognition. In: Proceedings of the 26th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01), Salt Lake City, pp. 309–312 (2001)
11. Whittmore, J., Fisher, S.: Speech during Sustained Operations. Speech Communications 20, 55–70 (1996)
12. Clary, G., Hansen, J.H.L.: A Novel Speech Recognizer for Keyword Spotting. In: Proceedings of the International Conference of Spoken Language Processing (ICSLP '02), Alberta, vol. 1, pp. 13–16 (1992)
13. Hansen, J.H.L., Bou-Ghazale, S.E.: Duration and Spectral Based Stress Token Generation for Keyword Recognition under Hidden Markov Models. IEEE Transactions on Speech & Audio Processing 3(5), 415–421 (1995)

14. Junqua, J.C.: The Lombard Reflex and its Role on Human Listeners and Automatic Speech Recognition. Journal of the Acoustic Society of America 93(1), 510–524 (1993)
15. Junqua, J.C.: The Influence of Acoustics on Speech Production: a Noise-Induced Stress Phenomenon known as the Lombard Effect. Speech Communication 20, 13–22 (1996)
16. Hicks, J.W., Hollien, H.: The Reflection of Stress in Voice-1: Understanding the Basic Correlates. In: Proceedings of the 1991 Carnahan Conference on Crime Countermeasures, pp. 189–195 (1981)
17. Hansen, J.H.L., Swail, C., South, A.J., Moore, R.K., Steeneken, H., Cupples, E.J., Anderson, T., Vloeberghs, C.R.A., Trancoso, I., Verlinde, P.: The Impact of Speech Under 'Stress' on Military Speech Technology. In: NATO RTO-TR-10, AC/323(IST)TP/5 IST/TG-01 (2000)
18. Murray, I.R., Baber, C., South, A.: Towards a Definition and Working Model of Stress and its Effects on Speech. Speech Communication 20, 3–12 (1996)
19. Goldberger, L., Breznitz, S.: Handbook of Stress: Theoretical and Clinical Aspects. Free Press, MacMilliam Pub., New York (1982)
20. Schreuder, M.J.: Prosodic Processes in Language and Music. PhD thesis, University of Groningen (2006)
21. Hansen, J.H.L.: Evaluation of Acoustic Correlates of Speech Under Stress for Robust Speech Recognition. In: IEEE Proceedings of the 15th Northeast Bioengineering Conference, Boston, pp. 31–32 (1989)
22. Paul, D.B.: A Speaker-Stress Resistant HMM Isolated Word Recognizer. In: Proceedings of the 12th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '87), Dallas, pp. 713–716 (1987)
23. Pickett, J.M.: The Sound of Speech Communication. University Park Press, Baltimore (1980)
24. Williams, C.E., Stevens, K.N.: Emotions and Speech: Some Acoustic Correlates. Journal of the Acoustic Society of America 52(4), 1238–1250 (1972)
25. Hansen, J.H.L.: Analysis and Compensation of Speech under Stress and Noise for Environmental Robustness in Speech Recognition. Speech Communications, Special Issue on Speech Under Stress 20(2), 151–170 (1996)
26. Van Santen, J.: Prosodic modeling in Text-to-Speech Synthesis. In: Proceedings of the European Conference on Speech Communication and Technology (Eurospeech '97), Rhodes, Greece, pp. 19–28 (1997)
27. Hansen, J.H.L.: Adaptive Source Generator Compensation and Enhancement for Speech Recognition in Noisy Stressful Environments. In: Proceedings of the 18th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '93), Minn., pp. 95–98 (1993)
28. Hecker, M.H.L., Stevens, K.N., von Bismark, G., Williams, C.E.: Manifestations of Task Induced Stress in the Acoustic Speech Signal. Journal of the Acoustic Society of America 44, 993–1001 (1968)
29. Hansen, J.H.L., Cairns, D.A.: ICARUS: Source Generator based Real-Time Recognition of Speech in Noisy Stressful and Lombard Effect Environments. Speech Communications 16(4), 391–422 (1995)
30. Hansen, J.H.L., Womack, B.: Feature Analysis and Neural Network based Classification of Speech under Stress. IEEE Transactions on Speech & Audio Processing 4(4), 307–313 (1996)
31. Womack, B.D., Hansen, J.H.L.: Classification of Speech Under Stress using Target Driven Features. Speech Communication, Special Issue on Speech Under Stress 20(1), 131–150 (1996)

32. Bou-Ghazale, S.E., Hansen, J.H.L.: Stressed Speech Synthesis Based on a Modified CELP Vocoder Framework. Speech Communications: Special Issue on Speech Under Stress 20(2), 93–110 (1996)
33. Hansen, J.H.L.: Morphological Constrained Enhancement with Adaptive Cepstral Compensation (MCE-ACC) for Speech Recognition in Noise and Lombard Effect. IEEE Transactions on Speech & Audio Proc (SPECIAL ISSUE: Robust Speech Recognition) 2(4), 598–614 (1994)
34. Hansen, J.H.L., Bria, O.N.: Lombard Effect Compensation for Robust Automatic Speech Recognition in Noise. In: Proceedings of the International Conference on Spoken Language Processing (ICLSP '90), Kobe, Japan, pp. 1125–1128 (1990)
35. Yapanel, U.H., Hansen, J.H.L.: A New Perspective on Feature Extraction for Robust In-Vehicle Speech Recognition. In: Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech '03), Geneva, Switzerland, pp. 1281–1284 (2003)
36. Bou-Ghazale, S.E., Hansen, J.H.L.: A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress. IEEE Transactions on Speech & Audio Processing 8(4), 429–442 (2000)
37. Hansen, J.H.L., Clements, M.A.: Constrained Iterative Speech Enhancement with Application to Speech Recognition. IEEE Transactions on Signal Processing 39(4), 795–805 (1991)
38. Clary, G., Hansen, J.H.L.: Feature Enhancement for Multi-layer Perceptron and Semi-Continuous Hidden Markov Model Based Classifiers using Neural Networks. In: Neural and Stochastic Methods in Image and Signal Processing, Proceedings of the SPIE, vol. 1766, pp. 529–540 (1992)
39. Cestaro, V.L.: A Comparison between Decision Accuracy Rates obtained using the Polygraph Instrument and Computer Voice Stress Analyzer (CVSA) in the absence of Jeopardy. Technical report, DOD Polygraph Inst. (1995)
40. Eriksson, A., Drygajlo, A.: Forsensic Speech Science. In: Tutorial, 9th European Conference on Speech Communication and Technology (Interspeech 05 - Eurospeech) (2005)
41. Zhou, G.: Nonlinear Speech Analysis and Acoustic Model Adaptation with Applications to Stress Classification and Speech Recognition. PhD thesis, Dept. of Electrical and Computer Eng., Duke University (1999)
42. Zhou, G., Hansen, J.H.L., Kaiser, J.: Linear and Nonlinear Speech Feature Analysis for Stress Classification. In: Proceedings of the International Conference on Spoken Language Processing (ICLSP '98), Sydney, Australia, vol. 3, pp. 883–886 (1998)
43. Zhou, G., Hansen, J.H.L., Kaiser, J.F.: Classification of Speech under Stress Based on Features Derived from the Nonlinear Teager Energy Operator. In: Proceedings of the 23th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98), Seattle, pp. 549–552 (1998)
44. Womack, B.D., Hansen, J.H.L.: N-Channel Hidden Markov Models for Combined Stress Speech Classification and Recognition. IEEE Transactions on Speech and Audio Processing 7(6), 668–677 (1999)
45. Kaiser, J.F.: Some Observations on Vocal Tract Operation from a Fluid Flow Point of View. In: Titze, I.R., Scherer, R.C. (eds.) Vocal Fold Physiology: Biomechanics, Acoustics, and Phonatory Control. Denver Center for the Performing Arts, Denver, pp. 358–386 (1983)
46. Teager, H.M.: Some Observations on Oral Air Flow during Phonation. IEEE Transactions Acoustic, Speech, Signal Processing 28(5), 599–601 (1980)
47. Teager, H.M., Teager, S.M.: A Phenomenological Model for Vowel Production in the Vocal Tract. In: Speech Science: Recent Advances, pp. 72–100 (1982)

48. Teager, H.M., Teager, S.: Evidence for Nonlinear Production Mechanisms in the Vocal Tract. In: NATO Advanced Study Inst. On Speech Production and Speech Modeling, Bonas, France, vol. 55, pp. 241–261. Kluwer Academic Publishers, Boston (1989)
49. Thomas, T.J.: A Finite Element Model of Fluid Flow in the Vocal Tract. Computer Speech Language 1, 131–151 (1986)
50. Hansen, J.H.L., Gavidia-Ceballos, L., Kaiser, J.F.: A Nonlinear based Speech Feature Analysis Method with Application to Vocal Fold Pathology Assessment. IEEE Transactions on Biomedical Engineering 45(3), 300–313 (1998)
51. Zhou, G., Hansen, J.H.L., Kaiser, J.F.: Nonlinear Feature Based Classification of Speech under Stress. IEEE Transactions on Speech & Audio Processing 9, 201–216 (2001)
52. Rahurkar, M., Hansen, J.H.L., Meyerhoff, J., Saviolakis, G., Koenig, M.: Frequency Band Analysis for Stress Detection Using a Teager Energy Operator Based Feature. In: Proceedings of the International Conference of Spoken Language Processing (ICSLP '02), Denver, vol. 3, pp. 2021–2024 (2002)
53. Ruzanski, E., Hansen, J.H.L., Meyerhoff, J., et al.: Stress Level Classification of Speech using Euclidean Distance Metrics in a Novel Hybrid Multi-Dimensional Feature Space. In: Proceedings of the 31st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06), Toulouse, vol. 1, pp. I–425–I–428 (2006)
54. Bou-Ghazale, S.E.: Analysis, Modeling, and Perturbation of Speech Under Stress with Applications to Synthesis and Recognition. PhD thesis, Robust Speech Processing Laboratory, Duke Univ. Dept. of Electrical Engineering (1996)
55. Bou-Ghazale, S.E., Hansen, J.H.L.: Stress Perturbation of Neutral Speech for Synthesis based on Hidden Markov Models. IEEE Transactions on Speech & Audio Processing 6(3), 201–216 (1998)
56. Cahn, J.: The Generation of Affect in Synthesized Speech. Journal of the American Voice I/O Society 8, 1–19 (1990)
57. Hansen, J.H.L., Clements, M.A.: Evaluation of Speech under Stress and Emotional Conditions. 82(S1), 7–8 (1987)
58. Murray, I.R., Arnott, J.L.: Implementation and Testing of a System for Producing Emotion-by-Rule in Synthetic Speech. Speech Communication 16, 369–390 (1995)
59. Murray, I.R., Arnott, J.L.: Synthesizing Emotions in Speech: is it time to get excited? In: Proceedings of the 4th International Conference on Spoken Language Processing (ICLSP '96), vol. 3, pp. 1816–1819. Philadelphia (1996)
60. Black, A.: Multilingual Speech Synthesis. In: Schultz, T., Kirchhoff, K. (eds.) Multilingual Speech Processing. Elsevier, Academic Press (2006)
61. Picard, R.W., Klein, J.: Computers that Recognize and Respond to User Emotion: Theoretical and Practical Implications. Interacting with Computers 14(2), 141–169 (2002)
62. Sproat, R. (ed.): Multilingual Text-to-Speech Synthesis: The Bell Labs Approach. Kluwer Academic Publishers, Boston (1997)
63. Van Santen, J., Kain, A., Klabbers, E.: Synthesis by Recombination of Segmental and Prosodic information. In: Proceedings of the International Conference on Speech Prosody, Japan, pp. 409–412 (2004)
64. Bachrach, A.J.: Speech and its Potential for Stress Monitoring: Monitoring Vital Signs in the Divers. Technical report, Naval Medical Research Institute (1979)
65. Chen, Y.: Cepstral Domain Talker Stress Compensation for Robust Speech Recognition. IEEE Transactions on Acoustic Speech Signal Process. 36, 433–439 (1988)

66. Darby, J.K.: Speech Evaluation in Psychiatry. Grune and Stratton, New York (1981)
67. Flack, M.: Flying Stress. Medical Research Committee, London (1918)
68. Hansen, J.H.L.: Analysis and Compensation of Noisy Stressful Speech for Environmental Robustness in Speech Recognition (invited tutorial). In: NATO-ESCA Proc. Inter. Tutorial & Research Workshop on Speech Under Stress, Lisbon, Portugal, pp. 91–98 (1995)
69. Hansen, J.H.L., Bou-Ghazale, S.E.: Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database. In: Proceedings of the European Conference on Speech Communication and Technology (Eurospeech '97), vol. 4, pp. 1743–1746. Rhodes, Greece (1997)
70. Hansen, J.H.L., Mammone, R., Young, S.: Editorial for the special issue: Robust Speech Recognition. IEEE transactions on Speech & Audio Processing 2(4), 549–550 (1994)
71. Hansen, J.H.L., Gavidia-Ceballos, L., Kaiser, J.F.: A Nonlinear based Speech Feature Analysis Method with Application to Vocal Fold Pathology Assessment. IEEE Transactions on Biomedical Engineering 45(3), 300–313 (1998)
72. Hollien, H., Hicks, J.W.: The Reflection of Stress in Voice-2: the Special Case of Psychological Stress Evaluators. In: Proceedings of the 1991 Carnahan Conference on Crime Countermeasures, pp. 196–197 (1991)
73. House, A.S.: On Vowel Duration in English. Journal of the Acoustic Society of America 33(9), 1174–1178 (1962)
74. Kuroda, I., Fujiwara, O., Okamura, N., Utsuki, N.: Method for Determining Pilot Stress Through Analysis of Voice Communications. In: Aviation, Space, and Environmental Medicine 528–533 (1976)
75. Kaiser, J.F.: Some Useful Properties of Teager's Energy operator. In: Proceedings of the 18th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '93), Minn., vol. 3, pp. 149–152 (1993)
76. Kaiser, J.F.: On a Simple Algorithm to Calculate the Energy of a Signal. In: Proceedings of the 15th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '90), Albuquerque, New Mexico, pp. 381–384 (1990)
77. McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M., Stroeve, S.: Approaching Automatic Recognition of Emotion from Voice: A rough Benchmark. In: Proceedings of the ISCA Workshop on Speech and Emotion, Belfast (2000)
78. Malkin, F.J., Christ, K.A.: Human Factors Engineering Assessment of Voice Technology for the Light Helicopter Family. Technical Report I-20, U. S. Armu Human Engineering Lab. (June 1985)
79. Maragos, P., Kaiser, J.F., Quatieri, T.F.: On Amplitude and Frequency Demodulation using Energy Operators. IEEE Transactions on Signal Processing 41, 1532–1550 (1993)
80. Poock, G.K., Armstrong, J.W.: Effect of Operator Mental Loading on Voice Recognition System Performance. Technical report, Naval Postgraduate School (1981)
81. Poock, G.K., Armstrong, J.W.: Effect of Task Duration on Voice Recognition System Performance. Technical report, Naval Postgraduate School (September 1981)
82. Schreuder, M., Eerten, L.v., Gilbers, D.: Music as a Method of Identifying Emotional Speech. In: Proceedings of the Workshop on Corpora for Research on Emotion and Affect (LRE '06), Genua, Italy, pp. 55–59 (2006)
83. Simonov, P.V., Frolov, M.V.: Analysis of the Human Voice as a Method of Controlling Emotional State: Achievements and Goals. Aviation, Space, and Environmental Sciences 23–25 (1977)

84. Streeter, L.A., MacDonald, N.H., Apple, W., Krauss, R.M., Galotti, K.M.: Acoustic and Perceptual Indicators of Emotional Stress. Journal of the Acoustic Society of America 73(3), 917–928 (1988)
85. Varadarajan, V., Hansen, J.H.L., Ikeno, A.: UT-SCOPE - A corpus for Speech under Cognitive/Physical Task Stress and Emotion. In: Workshop on Corpora for Research on Emotion and Affect (LREC '06), pp. 72–75 (2006)
86. Varadarajan, V., Hansen, J.H.L.: Analysis of Lombard effect under Different types and levels of Noise with Application to In-set Speaker ID systems. In: Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech '06 –ICSLP), Pittsburgh (2006)
87. Womack, B., Hansen, J.H.L.: Robust Speech Recognition via Speaker Stress Classification. In: Proceedings of the 31th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06), Toulouse, vol. 1, pp. 53–56 (2006)
88. Yamada, T., Hashimoto, H., Tosa, N.: Pattern Recognition of Emotion with Neutral Network. In: Proc. 21st Inter. Conf. on Industrial Electronics, Control, and Instrumentation (IECON '95), vol. 1, pp. 183–187 (1995)
89. Yapanel, U.H., Dharanipragada, S.: Perceptual MVDR-based Cepstral Coefficients for Noise Robust Speech Recognition. In: Proceedings of the 28th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03), Hong-Kong (2003)