S6.4

Stress Compensation and Noise Reduction Algorithms for Robust Speech Recognition

John H. L. Hansen and Mark A. Clements
Department of Electrical Engineering
Duke University
Durham, North Carolina 27706

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

1 Abstract

The problem of speech recognition in noisy, stressful environments is addressed. The main contribution is the achievement of robust recognition in diverse environmental conditions through the formulation of a series of speech enhancement and stress compensation preprocessing algorithms. These preprocessors produce speech or recognition features less sensitive to varying factors caused by stress and noise. Recognition results from four recognition scenarios based on enhancement and stress compensation preprocessing are reported. Neutral, stressful, noisy neutral, and noisy stressful speech styles are considered. Noise reduction is based on constrained iterative speech enhancement [1,2]. Stress compensation algorithms are based on formant location, bandwidth, and intensity. Enhancement preprocessing increases recognition by +34% for neutral speech, 18% for stressed speech. Combined stress compensation, speech enhancement preprocessing increases recognition rates by an average +27% (e.g., +43% loudly spoken speech, +42% speech spoken under Lombard effect). As a result, combined speech enhancement stress compensation preprocessing has been shown to be extremely effective in reducing the effects caused by stress and noise for robust automatic recognition.

2 Introduction

Previous studies in speech recognition have largely been directed at issues such as speaker restrictions, type of speech, and vocabulary size. There has been great interest, but limited progress in addressing the issue of diverse environmental conditions for speech recognition. This is due in part to the fact that past approaches such as dynamic time warping or hidden Markov modeling (HMM) have largely been applied in noise free tranquil environments. Studies have shown that recognition accuracy is severely reduced when speech is uttered in a noisy, stressful environment. If recognition is to be successful in such diverse environments, (e.g., pilots in aircraft cockpits, wheelchair control for the disabled, factory use for assembly lines), changing environmental conditions such as noise and stress must be taken into account.

The direction taken in this research has been the development of robust enhancement and stress compensation preprocessors. These preprocessors take advantage of past recognition techniques formulated in noise free tranquil environments by producing speech or recognition features which are less sensitive to varying factors such as stress and noise. The overall system configuration, illustrated in Figure 1, indicates three factors which affect speech entering the recognition system. First, background noise will have a degrading effect on the speech signal. Second, since the speaker is able to hear the background noise, he may alter his speech characteristics in an effort to increase communication efficiency over the noisy medium (i.e., the Lombard effect). Lastly, the performance of any secondary task may also affect characteristics of an operator's speech production system.

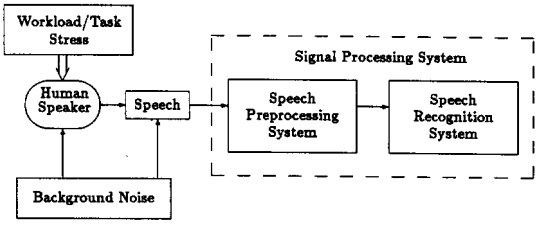


Figure 1: System environment for recognition of speech under noisy, stressful conditions.

Formulation of a solution requires the achievement of two goals. The first is to improve performance of recognition algorithms in noisy environments. A new set of constrained iterative speech enhancement algorithms were formulated for this purpose (ICASSP-87 [1], ICASSP-88 [2]), and function as enhancement preprocessors to reduce background noise prior to recognition. Section 3 discusses the enhancement algorithms used in this evaluation. The second goal is to improve recognition capabilities of speech produced under stressful conditions. To accomplish this, speech parameters most affected by environmental conditions must be identified. Section 4 summarizes results from a comprehensive investigation of speech under stress. Stress in this context refers to the result of factors which act on the speaker from environmental conditions (e.g., workload stress, background noise as in the Lombard effect, etc). This evaluation motivated the formulation of stress compensation preprocessing algorithms presented in Section 5. The final goal of robust recognition in noisy stressful environments is addressed in Section 6.

3 Constrained Iterative Enhancement

The set of speech enhancement algorithms under consideration were previously developed for improving both speech quality and all-pole speech parameter estimation [1,2]. The algorithms are based on sequential two step maximum a posteriori (MAP) estimation of the all-pole speech parameters $\vec{a}$ and noise free speech waveform $\vec{S}_O$. In order to improve parameter estimation, reduce frame to frame pole jitter across time, and provide a convenient and consistent terminating criterion, a variety of spectral constraints were introduced between MAP estimation steps. These constraints are applied based on the presence of perceptually important speech characteristics found during the enhancement procedure. The enhancement algorithms impose spectral constraints on all-pole parameters $\hat{\vec{a}}_i$ across time (inter-frame) and iterations (intra-frame) which ensure that; i) the all-pole speech model is stable, ii) it possess speech-like characteristics (e.g., poles are not too close to the unit circle causing narrow band-

CH2673-2/89/0000-0266 $1.00 © 1989 IEEE

widths), and iii) the vocal tract characteristics do not vary wildly from frame to frame when speech is present. Due to the imposed constraints, improved estimates of $\hat{\vec{a}}_{i+1}$ result. In order to increase numerical accuracy, reduce computational requirements, and eliminate inconsistencies in pole ordering across frames, the line spectral pair (LSP) transformation was used to implement most of the constraint requirements. This method allowed constraints to be efficiently applied to speech model pole movements across time so that formants lay along smooth tracks. In addition, constraints are also easily applied across iterations on a frame-by-frame basis. Figure 2 illustrates the enhancement framework.

These algorithms were shown to be preferable to existing techniques in several respects. First, results based on objective speech quality measures show that the current systems result in substantially improved speech quality and LPC parameter estimation over past techniques. Second, the enhancement algorithms have been shown to perform well on non-stationary colored noise. Third, the current algorithms have been shown to possess a much more consistent terminating criterion. Specifically, the optimum terminating iteration was shown to be consistent over all speech sound classes, and virtually all tested SNR's, giving a simple procedure for termination of the algorithm. Finally, the constrained algorithms have been shown to be superior in estimating LPC parameters as measured by distance measures normally used for LPC parameters (log-area-ratios, Itakura-Saito distances, etc.) Estimation of the vocal-tract response is also substantially better in the current systems. This represents an important feature in preprocessing for robust recognition. Detailed discussions of these algorithms can be found in [1,4,5].

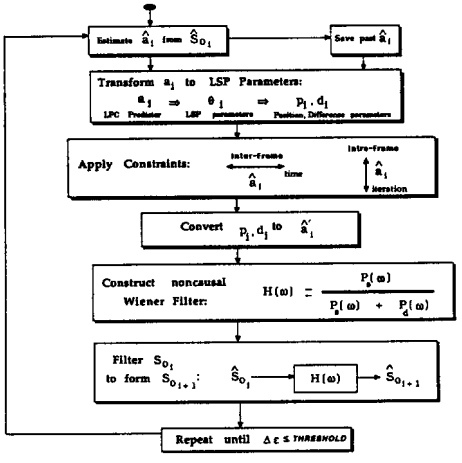


Figure 2: Framework for the class of constrained enhancement algorithms.

4 Analysis of Speech Under Stress

The next step is to identify which speech parameters are most affected by environmental conditions such as stress or noise. Previous research directed at this problem has generally been limited in scope, often suffering from one to five problems. These include: i) limited speaker populations, ii) sparse vocabularies, iii) qualitative results with little statistical confirmation, iv) limited numbers and types of speech parameters considered, and v) analysis based on simulated or actual conditions with little confirmation between the two. In order to address these issues, a comprehensive investigation was performed to reveal new and statistically reliable acoustic correlates of speech under stress [3,5,6]. Careful planning was necessary in formulating and collecting a speech under stress data base for analysis. Table 1 illustrates the five domains of the data base. A total of 32 speakers were employed to generate in excess of 16,000 utterances. The data base was partitioned into two areas for analysis, i) simulated stress or emotional conditions and ii), actual stressed conditions or effects caused by noise. Speech parameter domains considered in the analysis include characteristics of pitch, glottal source spectrum, duration, intensity, and vocal-tract shaping (approximately 200 speech parameters were considered). Extensive statistical evaluations were performed to identify the significance of variations in average, variance, and distribution of each parameter. Results show that characteristics of the pitch period represent some of the best stress discriminating parameters. Glottal source characteristics (e.g., spectral tilt, average spectral energy), resulted in wide variations across stress styles. Finally, first and second formant location and bandwidth parameters, along with the variability of these, were very reliable stress indicators, especially for vowels. Further discussion can be found in [3,5,6].

Speech Under Stress Data Base Georgia Institute of Technology School of Electrical Engineering				
Domain	*Type of Stress or Emotion*	*Number of Speakers*	*Number of Utterances*	*Source*
Psychiatric Analysis	Depression Fear, Anger Anxiety	6 Female, 2 Male	600 (present)	Emory University School of Medicine Department of Psychiatry
Talking Styles	slow, fast soft, loud angry, clear question	All Male 3 General 3 New York 3 Boston	5820 (total)	Lincoln Labs Boston, Mass. 35 aircraft communication words
Single Tracking Task	*Workload* (moderate-C50) (high-C70) Lombard	All Male	1890 (total)	Lincoln Labs Calibrated Workload Tracking Task
Dual Tracking Task	*Workload* (moderate) (high)	4 Female 4 Male	4320 (total)	Georgia Tech Acquisition tracking Compensatory tracking
Subject Motion-Fear Tasks	G-force Lombard,noise anxiety,fear	3 Female 4 Male	400 (total)	Georgia Tech Controlled Motion Noisy Environment

Table 1: The Georgia Tech Speech under Stress Data Base.

5 Stress Compensation Algorithms

The motivation for the analysis of speech under stress was to uncover those acoustic correlates which vary under stressful conditions. Variation in these parameters may suggest a possible explanation for adverse recognition performance in diverse environments. The previous investigation explored areas of speech production which traditionally have not been associated with present day recognition algorithms. The reasons for this are twofold. First, it may be possible to improve existing recognition algorithms by allowing preprocessors to reduce or eliminate parameters which are affected by stress. Although the effects of some parameters (i.e., pitch, duration, intensity) are somewhat mitigated in many recognition procedures, severe variations do adversely affect recognition performance (e.g., HMM recognizers have certain "time constants" which tolerate only a limited degree of duration variability). Other stress analysis parameters such as characteristics of glottal source spectrum (spectral tilt, energy distribution) and vocal-tract (formant center frequencies, bandwidths, spectral tilt, and the variability of these) have direct consequences in recognition performance. Second, other stress relaying parameters not used for recognition, may be used to reliably identify when an utterance is under stress so that appropriate stress compensation can be employed.

A set of stress compensation algorithms were formulated based on results from three stress analysis domains. These approaches assume the stress condition (e.g., loud, angry, clear, etc.) to have already been identified. The algorithms are based on obtaining a table of compensation factors for all phonemes, for each stress condition. The three possible processing steps include: i) compensate for average formant location (F1,F2,F3,F4), ii) compensate for average formant bandwidth (B1,B2,B3,B4), iii) and compensate for overall word intensity. In order to calculate formant location and bandwidth values, root solving and pole ordering of the LPC polynomial for each speech frame was performed. To reduce the variance of average formant location and bandwidth estimates, a smoothing operation was performed prior to calculation of average formant values. This served to improve the estimation of average values by reducing the effects of outlying values caused by misclassification during ordering. Formant compensation factors were obtained by taking the ratio of average formant values between neutral and stressed conditions. Parametric and non-parametric statistical tests were used to verify significance of variation in formant characteristics. Table 2 presents sample compensation factors for average formant location and bandwidth used for the angry stressed condition. As an example, consider the compensation for the vowel /e/. The term F1 (= 0.63), was used to decrease all first formant locations for phoneme /e/ under angry conditions. The average first formant location will then have the same average value as that found in neutral conditions. This process was repeated for each formant location and bandwidth. A similar table was obtained for each of the ten stress conditions. Once the compensation tables are known, preprocessing stress compensation algorithms were implemented. Two additional points are necessary. First, unlike the fully automated constrained speech enhancement algorithms summarized in Section 2, the stress compensation algorithms require both knowledge of the type of stress and phoneme boundaries in order to apply compensation factors. Further research is underway to incorporate general stress compensation within the constrained enhancement algorithms, thereby removing this requirement. Second, as demonstrated in [1], such compensation schemes could easily be extended to LSP parameters, and therefore integrated within the speech enhancement algorithms. This particular approach was chosen since computational requirements were not at issue, and that shifts in formant location and bandwidth gave a more intuitive feel for how the vocal-tract spectrum was being adjusted.

		FORMANT COMPENSATION FACTORS							
Category	*Phoneme*	F_1	F_2	F_3	F_4	B_1	B_2	B_3	B_4
CONSONANTS									
nasals	/N/	.74	1.04	.99	1.01	.71	1.32	1.41	1.39
stops									
voiced:	/D/	.92	.96	1.02	.98	1.26	.81	.65	.93
unvoiced:	/K/	.96	.99	1.02	1.02	.84	.53	.87	.80
whisper	/H/	.85	.95	.95	1.04	.97	.66	1.25	.67
affricates	/TSH/	1.27	.93	1.02	1.01	.96	1.06	1.02	1.08
fricatives									
voiced:	/TH/	1.24	1.05	1.05	1.02	1.03	1.10	.76	1.01
unvoiced:	/S/	.88	.93	1.01	1.07	1.28	.72	.96	.93
VOWELS									
front:	/IY/	.70	.93	.96	.98	.60	1.38	1.09	.92
mid:	/e/	.63	.92	.97	.95	.61	.96	.68	1.19
DIPHTHONGS									
	/oU/	.60	.94	1.00	1.01	.96	1.36	1.49	1.13
SEMIVOWELS									
liquid:	/W/	1.02	1.51	1.05	1.00	1.26	1.35	1.12	.93

Table 2: Sample compensation factors for average formant location (F1,F2,F3,F4) and bandwidth (B1,B2,B3,B4) for angry spoken speech.

6 Recognition Framework & Results

Advances made in the analysis of speech under stress and speech enhancement domains are joined to address the final goal of recognition in noisy stressful environments.

A fairly standard, isolated-word, discrete-observation hidden Markov model recognition system was used for evaluation. This system was LPC based and had no embellishments. In all experiments, a five state, left-to-right model was used. System dictionary consisted of twenty highly confusable words from the second and third domains of the speech under stress data base. These words are also used by Texas Instruments and Lincoln Labs to evaluate recognition systems. Subsets include {go, oh, no}, {six, fix}, and {wide, white}. Thirty-two examples of each word were used in the evaluation, six neutral examples for training, six neutral examples for recognition, and two examples for each of the ten stressed speaking styles (i.e., soft, loud, etc.) for recognition (i.e., all tests fully open employing a neutral trained HMM system). A vector quantizer was used to generate a 64 state codebook using two minutes of noise free, neutral training data. The twenty models employed by the HMM recognizer were trained using the forward-backward algorithm.

Figure 3 illustrates the recognition scenarios in the evaluation. Results from each are summarized in Figure 4. The first four evaluations establish baseline recognition scores for neutral, stressful, noisy neutral, and noisy stressful speech conditions. The recognition rate of noise free neutral speech (88%) confirms the confusability of the chosen vocabulary. Independent evaluations of this system with distinct vocabularies resulted in recognition rates of 100% [5]. Baseline scores indicate that stress, with and without background noise, has a profound effect on recognition performance. Recognition rates dropped by an average 31% for stressful speech, with an additional 19% for noisy stressful speech. Thus indicating that recognition degrades rapidly whether a speaker is under stress, in noise, or a combination.

The fifth recognition scenario employed enhancement preprocessing of noisy neutral speech. In ICASSP-88[2], the constrained enhancement algorithms were shown to be superior to implementations of past enhancement techniques (e.g., spectral subtraction, noncausal Wiener filtering) in preprocessing for recognition of noisy neutral speech. Therefore, only the constrained enhancement techniques are considered here. The constrained enhancement algorithm used (FF-LSP:T,Auto:I) was based on fixed-frame constraints applied across time, and constraints applied to autocorrelation lags across iterations (see [1,5] for further discussion). The noise degradation was additive white Gaussian, with SNR's

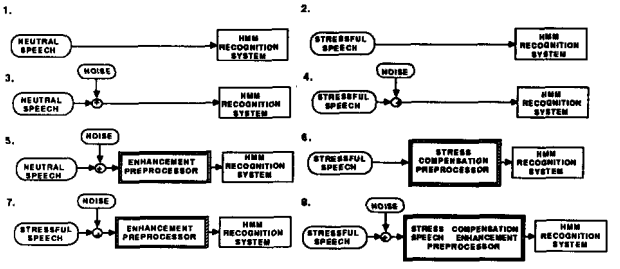
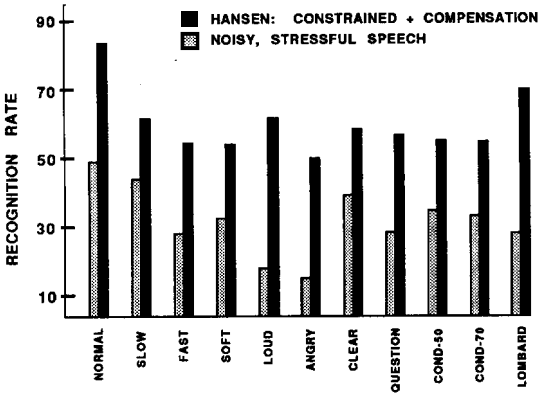


Figure 3: Robust speech recognition scenarios employing enhancement and/or stress compensation preprocessing.

determined over entire utterances. A 34% increase in recognition was observed for enhanced neutral speech. For the sixth recognition scenario, the same enhancement preprocessing was employed for noisy, stressful speech. Recognition rates significantly increased for all types of stress (an average +17.8%). It should also be noted that SNR's in low energy consonantal portions which discriminate confusable pairs (e.g., "go - oh - no") may well be 20 dB lower then global SNR measurements. The enhancement preprocessors are therefore successful in reducing background noise as well as reducing some vocal-tract variations caused by stress.

Next, stress compensation preprocessing of noise free stressful speech was considered. Three stress compensation algorithms were evaluated, i) average formant location compensation (FL), ii) average formant bandwidth compensation (FB), iii) combined formant location and bandwidth compensation (FL+FB). All compensators included intensity compensation. Figure 4 presents results from these evaluations. Collectively, nine of the ten stressed conditions benefited from stress compensation. FL+FB is preferable for varying vocal effort (soft,loud) and angry speech (half of all recognition errors were eliminated). Stress compensation did not improve recognition performance for the clear speaking style, thereby suggesting that other stress factors (beside formant location and bandwidth) should be considered. Finally, for speech under the Lombard effect, FB compensation provided the best recognition improvement (+13%). Overall recognition performance was consistent across varying stress styles, indicating the success in reducing effects caused by stress.

RECOGNITION EMPLOYING STRESS COMPENSATION and SPEECH ENHANCEMENT



Condition	N	S	F	So	L	A	C	Q	C50	C70	Lom
Noise free,Stressful	89%	60%	65%	48%	50%	20%	66%	75%	53%	53%	65%
STRESS COMPENSATION RECOGNITION RESULTS											
Compensator FL		65%	58%	53%	65%	44%	55%	50%	70%	60%	53%
Compensator FB		75%	70%	58%	50%	28%	50%	60%	75%	65%	74%
Compensator FL+FB		68%	58%	58%	68%	65%	53%	63%	77%	74%	67%
ENHANCEMENT & STRESS COMPENSATION RECOGNITION RESULTS †											
Condition	N	S	F	So	L	A	C	Q	C50	C70	Lom
Noisy,Stressful	49%	45%	28%	33%	18%	15%	40%	28%	35%	33%	28%
FF-LSP:T,Auto:1	83%	67%	53%	43%	36%	28%	58%	65%	58%	53%	38%
plus Compensator FL		50%	47%	53%	50%	45%	47%	35%	50%	55%	55%
plus Compensator FB		56%	37%	53%	56%	25%	56%	68%	55%	55%	70%
plus Compensator FL+FB		61%	53%	53%	61%	50%	53%	56%	55%	55%	65%

Speaking Styles Key:
N — neutral; F — fast; L — loud; C — clear; C50 — Task Condition 50 %
Sl — slow; So — soft; A — angry; Q — question; C70 — Task Condition 70 %; Lom — Lombard effect

Figure 4: Recognition results of neutral, stressed, noisy neutral, and noisy stressed speech with, and without preprocessing. The graph illustrates the best recognition improvement employing combined stress compensation speech enhancement preprocessing. †Additive white Gaussian noise, SNR = +30dB

The final recognition evaluation combined enhancement and stress compensation preprocessing. In half of the noisy stressful conditions, compensation did not appreciably raise recognition rates over enhancement preprocessing alone, thus suggesting that either enhancement preprocessing has the performed necessary stress compensation, or that other forms of compensation are required. Improvement was observed in several key stress styles (e.g., loud, angry, Lombard). Increased recognition ranged from +22% to +27% over enhancement preprocessing alone, and +35% to +43% over original noisy stressful speech. These results are encouraging, since recognition in loud, angry, and Lombard conditions are most closely associated with speech from actual noisy stressful environments (such as an aircraft cockpit). The graph in Figure 4 summarizes the best combined enhancement, stress compensation preprocessing results. A cursory inspection reveals consist recognition performance over varying noisy stress conditions, thereby indicating the effectiveness of preprocessing robust speech recognition.

7 Conclusions

The problem of speech recognition in noisy, stressful environments has been addressed in this paper. A series of speech enhancement and stress compensation preprocessing algorithms were formulation which produce speech or recognition features which are less sensitive to varying factors caused by stress and noise. Previous results have shown the constrained enhancement algorithms to improve recognition performance for neutral speech over past enhancement techniques for a wide range of SNR's. Enhancement preprocessing also results in marked increases in recognition under noisy stressful conditions. Stress compensation techniques (based on formant location, bandwidth, and intensity), have been shown to reduce the effects of stress present in changing vocal-tract characteristics, thereby improving recognition of noise free stressful speech. Finally, combined stress compensation, speech enhancement preprocessing increased recognition rates by an average +27% (e.g., +43% loudly spoken speech, +42% speech spoken under Lombard effect). In conclusion, combined speech enhancement and stress compensation preprocessing has been shown to be extremely effective in reducing the effects caused by stress and noise for robust automatic recognition.

This research sponsored in part by DoD.

References

[1] J.H.L. Hansen, M.A. Clements, " Iterative Speech Enhancement with Spectral Constraints," *Proc. 1987 IEEE ICASSP*, pp. 189-192, Dallas, TX, April 1987.

[2] J.H.L. Hansen, M.A. Clements, " Constrained Iterative Speech Enhancement with Application to Automatic Speech Recognition," *Proc. 1988 IEEE ICASSP*, pp. 561-564, New York, NY, April 1988.

[3] J.H.L. Hansen, M.A. Clements, " Evaluation of Speech under Stress and Emotional Conditions," *Proc. of the Acoustical Society of America*, 114th Meeting, Miami, Florida, Nov. 1987.

[4] J.H.L. Hansen, M.A. Clements, " Constrained Iterative Speech Enhancement, " *Trans. on Acoustics, Speech, and Signal Processing*, pp. 1-31, March 1988, in review.

[5] J.H.L. Hansen, " Analysis and Compensation of Stressed and Noisy Speech With Application to Robust Automatic Recognition," Ph.D. Thesis, Georgia Institute of Technology, 396 pages, July 1988.

[6] J.H.L. Hansen, "Evaluation of Acoustic Correlates of Speech Under Stress for Robust Speech Recognition." invited paper *Proc. Northeast Bioengineering Conference*, Boston, Mass., March 1989.