DURATION AND SPECTRAL BASED STRESS TOKEN GENERATION FOR HMM SPEECH RECOGNITION UNDER STRESS*

Sahar E. Bou-Ghazale and John H. L. Hansen

Digital Speech Processing Laboratory
Department of Electrical Engineering
Duke University, Box 90291
Durham, North Carolina 27708-0291

ABSTRACT

In this paper, we address the problem of isolated word recognition of speech under various stressed speaking conditions. The main objective is to formulate an alternate training algorithm for hidden Markov model recognition, which better characterizes actual speech production under stressed speaking styles such as slow, loud and Lombard effect, *without* the need for collecting such stressed speech data. The novel approach is to first construct a previously suggested source generator model of word production employing knowledge of the statistical nature of duration and spectral variation of speech under stress. This is used in turn to produce simulated stressed speech training tokens from neutral tokens, and thus replace neutral data used in the recognizer training phase. The token generation training method is shown to improve isolated word recognition by 8% for slow speaking style, 14% for loud speaking style, and 24% for speech under Lombard effect when compared to neutral trained isolated word recognition.

1. INTRODUCTION

Studies conducted on human speech production show that, depending on task workload, environmental conditions such as a high noise environment, as well as the speaker's physical and psychological state, a person's speech characteristics change with time. Examples of environmental settings which affect the manner of speech production include, i) task-workload (flying an aircraft, operating an automobile), (ii) normal intra speaker variability such as variable speaking style (e.g., speech spoken loud, soft, fast, slow, etc.), or (iii) speech spoken in noise (i.e., Lombard effect [9]). The variability introduced by a speaker under stress causes neutral token trained recognizers to fail [2, 3, 6, 8]. Unlike the human auditory system, which is capable of extracting this variability as additional perceptual information of the speaker (i.e., emotion, situational speaker state), typical recognition algorithms do not attempt to extract this information and and do not address such speaking conditions. It has been shown that recognition rates drop for a discrete observation hidden Markov model (HMM) system by 28% for slow speech, 25% for Lombard speech, and 38% for loud speech [6]. The goal of this study is to formulate a procedure for artificially generating stressed speech tokens for HMM training to achieve improved automatic recognition of speech spoken under stressful conditions. This is not intended to result in solving the problem of robust automatic speech recognition under stress, but instead to better understand the physiological based speech production variations under stress, and to determine the influence of including such knowledge within a HMM training procedure. Artificially simulated stressed tokens are neutral tokens for which speech production features such as duration and frequency content have been altered to statistically resemble stressed speech tokens. The proposed training procedure differs from a direct modification of the HMM word model, since direct modification may not fully portray the actual changes in speech production under stress. The stressed speaking conditions which constitute the focus of this study are slow, loud and Lombard effect.

*This work sponsored in part by National Science Foundation NSF-IRI-90-10536, and Naval Research and Development N66001-92-D-0092

2. PREVIOUS APPROACHES

In a previous study, Lippmann, Martin, and Paul proposed a multistyle training procedure for speech recognition in stress [8]. Though successful, this technique requires speakers to produce simulated speech tokens as they might perceive that stressed condition. Consequently, the speech tokens may fail to characterize the actual variability of speech in an actual stressed speaking environment. Hansen and Clements [6] proposed compensating for formant bandwidth and formant location in the recognition phase. This resulted in improved speech recognition unver noisy stressful conditions. However, this compensation required knowledge of phoneme boundaries and is computationally expensive. Alternatively, Chen [2] compensated for cepstral changes through transformations using neutral token training. This compensation is applied to word models where all speech sections within a word use the same compensation vector. Since stress does not affect all phonemes of a word equally[3], Hansen and Bria proposed a new approach for mel-cepstral compensation [7]. In their approach, each word is partitioned in the time domain into three broad speech classes (e.g., voiced/transitional/unvoiced speech sections). Similar speech classes are then grouped together and compensation is applied to each speech class separately for each

0-7803-1775-0/94 $3.00

isolated word model. Compensating for the effect of stress in the recognition phase was shown to improve speech recognition by 41.33%. However, such compensation can introduce delay in the recognition process since a separate compensator is integrated within each HMM word recognizer.

3. HMM-BASED STRESSED TOKEN GENERATION

The main objective here is to achieve reliable recognition performance under stressed speaking conditions without the requirement of having speakers produce simulated stress tokens. This is achieved by suggesting a training procedure which employs statistically generated stress speech tokens. The motivation for generating simulated stress tokens is due to (i) the inconvenience of collecting stress data for training from users, and (ii) the inaccuracy of human simulated stress tokens in representing actual speech under stress. A brief discussion of each processing phase is discussed below. Further details can be found in [1].

3.1. Source Generator Framework

As described in [4], the production of an isolated word can be described as a sequence of speech articulator movements to achieve desired vocal tract target shapes. This collection of speech articulator movements is represented by a sequence of source generators $\gamma_j; j = 1, 2, ...J$ in an F-dimensional feature space. Speech production of a word can be modeled as a sequence of movements from one source generator to another in this feature space. These movements represent a well defined path between source generators in the F-dimensional space. When a particular stress condition is introduced, deviations will result in this path. The analysis of these deviations will be used to model the effects of stress speaking styles on speech.

3.2. Source Generator Based Analysis

For the purpose of this study, a source generator sequence is assumed to represent detected phoneme-like speech classes partitioned into voiced/transitional/unvoiced (v/tr/uv) regions across time. The v/tr/uv classifier is based on the energy contour of an input utterance [5]. Here, a statistical study is conducted to develop models for duration and spectral content of stressed speech source generators. These models are later used to adapt the source generators duration and spectral features of neutral speech. A source generator is characterized by its duration $d^{(s)}_{i,j,k}$ and spectral content as represented by mel-cepstral parameters $C^{(s)}_l(i,j,k)$. Each source generator duration $d^{(s)}_{i,j,k}$ of a word k is assumed Gaussianly distributed with mean $\mu^{(s)}_{j,k}$, and variance $\sigma^{2(s)}_{j,k}$, and is characterized by the following equations:

$$f^{(s)}_{j,k}(d) = \frac{1}{\sqrt{2\pi}\sigma^{(s)}_{j,k}} \exp^{\frac{(d-\mu^{(s)}_{j,k})^2}{2\sigma^{2(s)}_{j,k}}} \quad (1)$$

$$\mu^{(s)}_{j,k} = \frac{1}{I}\sum_{i=1}^{I} d^{(s)}_{i,j,k} \quad (2)$$

Style	γ_1 (msec)		γ_3 (msec)		γ_5 (msec)	
	μ_1	σ_1	μ_3	σ_3	μ_5	σ_5
Neutral	47	22	376	42	46	34
Slow	98	91	760	214	125	96
Loud	75	44	439	81	65	21
Lombard	78	55	458	111	83	63

Table 1: Source generators durations for *zero* across different speaking styles.

$$\sigma^{2(s)}_{j,k} = \frac{1}{I-1}\sum_{i=1}^{I}(d^{(s)}_{i,j,k} - \mu_{j,k})^2. \quad (3)$$

where s spans the possible stressed speaking styles, k spans all possible keywords from a speaking style, i spans the domain of all the available tokens of a keyword collected from all speakers, j spans the number of source generators for that word, and I is the total number of tokens.

Next, stressed speech production variation within spectral content are modeled for artificial token generation. One mel-frequency cepstral coefficient (MFCC) vector $\hat{m}^{(s)}_{C_l}(j,k)$ is found for each source generator j of a word k across all available tokens. A mel-cepstral adaptation factor of a single source generator within a word is given by equation (4) below. The MFCC of individual source generators of a neutral word are scaled according to equation (5) to yield the adapted mel-cepstral vector. The spectral characteristics of word k spoken under style s are represented by the following equations:

$$\rho_{mean}(j,k)_l = \frac{\hat{m}^{(s=Str)}_{C_l}(j,k)}{\hat{m}^{(s=Ntr)}_{C_l}(j,k)} \quad (4)$$

$$C^{(s^*)}_l(j,k) = [\rho_{mean}(j,k)_l \times C^{(s=Ntr)}_l(j,k)]. \quad (5)$$

where l spans the number of extracted mel-cepstral coefficients per frame. Table 1 along with Fig. 1 ilustrate the duration and spectral characteristic changes from one source generator class to another within the same style, as well as the differences among similar source generators across different speaking styles.

3.3. Simulated Stress Token Generation

Generating simulated stress tokens is achieved by first adapting the individual source generator duration of neutral tokens, followed by controlled perturbation of the spectral characteristics of each duration modified neutral token (refer to the flow diagram in Fig. 2). This adaptation/perturbation is achieved by modeling the changes from each source generator class between neutral and each stressed speaking style, and using these models for controlled perturbation of neutral speech during HMM training. Duration adaptation is performed by (i) statistically determining the source generator durations of a typical stress token using the duration PDF's generated earlier, and (ii) adapting the source generator durations of the neutral token at the mel-cepstral parameterization stage. A

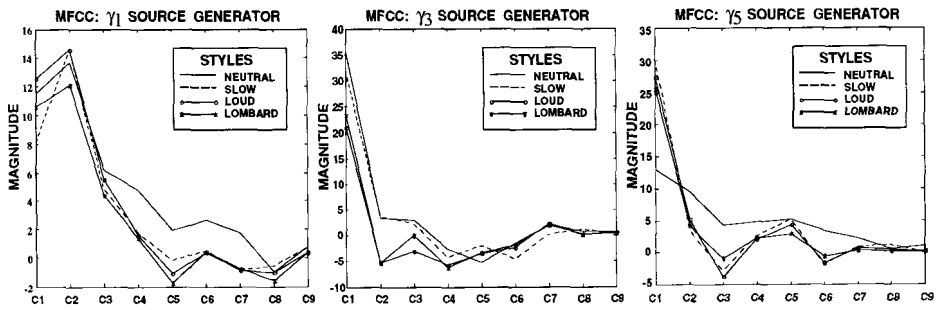


Figure 1: MFCC for source generators γ_1, γ_3, and γ_5 of the word *zero* across different speaking styles.

fixed-length, variable skip rate Hamming analysis window is used for extracting mel-cepstral coefficients. Depending on the stress dependent statistical durations, more observations are extracted from a source generator section and submitted to the recognizer if an increase in duration is required for a source generator, and less observations are extracted in case of a decrease. After extracting the required number of stressed duration based mel-cepstral vectors, each vector within a speech class j is perturbed using equation (5) for each mel-cepstral coefficient for stress condition s.

After information has been obtained from this statistical characterization, the speaker is no longer required to produce speech tokens under stress conditions. Subsequent training requires collecting only neutral tokens from the speaker, with application of the apriori estimated statistical model for that stress condition.

4. RECOGNITION RESULTS

The statistical studies conducted in this work are based on a previously collected data base, called *SUSAS(Speech Under Simulated and Actual Stress)* [3]. A discrete observation, 5-state left-to-right, isolated word HMM recognizer was used for all experiments. The stress word models were created using a total of 12 training tokens per word (6 neutral + 6 simulated stress). A 256 entry vector quantizer codebook was generated from a 35-word vocabulary spoken by one speaker under normal, slow, loud, and Lombard conditions [3].

The data used for this study consists of isolated words spoken by 9 male speakers under the three stressed speaking styles of *slow, loud,* and *Lombard.* The confusion matrices in Figure compare the performance of our system to the neutral trained system. The shade in each block is directly proportional to the recognition rate. A darker shade along the diagonal indicates improved performance. As can be seen, the proposed system consistently improves recognition performance. The best performance is achieved for the word "ten", for which recognition increased from 36% to 86%. When compared to neutral trained models in isolated word recognition, the simulated stress trained models as illustrated in Figure 4, improved recognition by 8% for slow speaking style, 14% for loud speaking style, and 24% for Lombard speech.

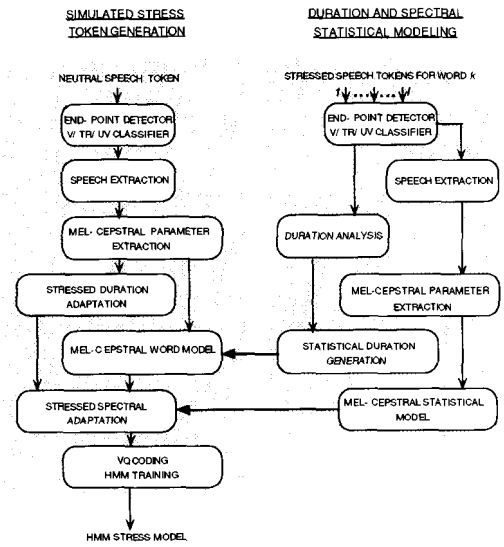


Figure 2: A block diagram representation of simulated stress tokens generation.

5. CONCLUSION

A new approach for generating simulated stressed training tokens has been presented and demonstrated for a discrete observation hidden Markov model recognizer in isolated speech scenario. The generated artificial stress tokens

can allow a user a controlled level of variability in the degree of stress for recognizer training. This provides a training algorithm which is capable of representing a wider range of stressed speech levels in speakers by adjusting source generator modeling features.

The simulated stress trained speech recognizer presented in this work improved overall recognition when tested on a sample data set from three stressed speaking styles. Duration and mel-cepstral parameter modification proved to be effective in improving stressed speech recognition. The proposed training method could potentially be used to model day-to-day speaker variability caused by task stress. Also, the results of this study suggest that training with generated stress tokens is a promising approach for achieving stress resistant and/or speaker-independent isolated word recognition in adverse (task-demanding) conditions.

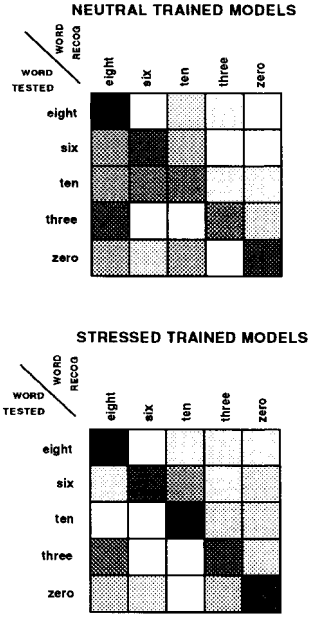
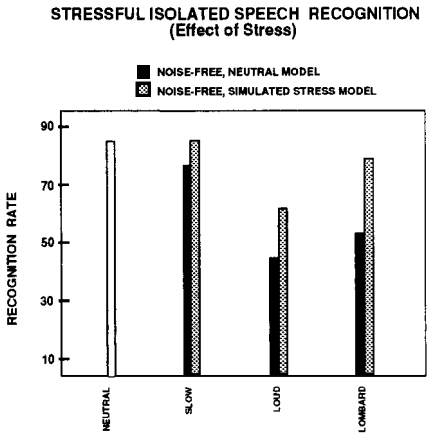


Figure 3: Performance of the neutral and simulated stress trained isolated word recognizers for stressed speech conditions.



ISOLATED WORD RECOGNITION			
Training Models With	*Testing Models With*		
	Slow	*Loud*	*Lombard*
Neutral speech	76%	44%	52%
Simulated stress speech	84%	58%	76%
Improvements	+8%	+14%	+24%

Figure 4: Isolated word recognition performance of the neutral trained models and the simulated stressed speech trained models when tested with stressed speech in noise-free conditions.

References

[1] Sahar E. Bou-Ghazale. " Duration and Spectral Based Stress Token Generation for Keyword Recognition Using Hidden Markov Models ". Master's thesis, Duke University, Department of Electrical Engineering, June 1993. Technical Report DSPL-93-11.

[2] Y. Chen. "Cepstral Domain Stress Compensation for Robust Speech Recognition". In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 717–720, Dallas, Texas, April 1987.

[3] J. H. L. Hansen. *"Analysis and Compensation of Stressed and Noisy Speech with Application to Robust Automatic Recognition"*. PhD thesis, Georgia Institute of Technology, Atlanta, Georgia, July 1988.

[4] J. H. L. Hansen. "Morphological Constrained Enhancement with Adaptive Cepstral Compensation for Speech Recognition in Noise and Lombard Effect". *Submitted to IEEE Transactions on Speech and Audio Processing*, October 1992. Tech. Report DSPL-92-5.

[5] J. H. L. Hansen and O. Bria. "Improved Automatic Speech Recognition in Noise and Lombard Effect". *EURASIP-92, The Sixth European Signal Processing Conference*, pages 403–406, August 1992.

[6] J. H. L. Hansen and M.A. Clements. "Stress Compensation and Noise Reduction Algorithms for Robust Speech Recognition". In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 266–269, Glasgow, Scotland, May 1989.

[7] J.H.L. Hansen and O.N. Bria. "Lombard Effect Compensation for Robust Automatic Speech Recognition in Noise ". In *ICSLP-90, Inter. Conf. on Spoken Language Processing*, pages 1125–1128, Kobe, Japan, November 1990.

[8] R. P. Lippmann, E. A. Martin, and D. B. Paul. "Multi-Style Training for Robust Isolated-Word Speech Recognition". In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 705–708, Dallas, Texas, April 1987.

[9] E. Lombard. "Le Signe de l'Elevation de la Voix". *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, 37:101–119, 1911.