A Novel Training Approach For Improving Speech Recognition Under Adverse Stressful Conditions

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

ABSTRACT

This paper presents a new training approach for improving recognition of speech under emotional and environmental stress. The proposed approach consists of training a speech recognizer with synthetically generated speech under each stress condition using stress perturbation models previously formulated in [4, 1]. The perturbation models were previously formulated to statistically model the parameter variations under angry, loud, and Lombard effect and were employed in an analysis-synthesis scheme for generating stressed synthetic speech from isolated neutral speech. In this paper, two training approaches employing the synthetically generated stressed speech are presented consisting of : speaker-independent, and speaker-adaptive training methods. Both approaches outperform neutral trained recognizers when tested with angry, loud, and Lombard effect speech.

1. INTRODUCTION

The variability introduced by a speaker under stress causes the performance of recognizers trained with neutral tokens to degrade. Training a model with speech produced under the same testing conditions can lead to improved recognition performance. However, such speech data is not always readily available for training. Several approaches have been previously proposed for improving stressed speech recognition. An approach referred to as *multi-style training* by Lippmann *et al.* [11] was considered for improved speaker-dependent recognition of stressed speech. This method required speakers to produce speech under simulated stressed speaking conditions and employed these multi-styles within the training procedure. A later study showed that multi-style training actually degrades performance if employed in a speaker-independent application [12]. Hansen and Clements [10] proposed compensating for formant bandwidth and formant location in the recognition phase. *Front-end feature modification* of stress speech in the recognition phase such that stress speech recognition parameters resemble that from neutral speech is another approach considered by Chen [6], Hansen and Bria [9], and Hansen [7] to improve recognition performance under stress. These methods all result in improved recognition performance. An alternative approach, referred to as the *token generation method*, altered both duration and mel-cepstral parameters of neutral training data to statistically resemble stressed speech data [2]. The token generation training method, which was tested on a limited vocabulary and was text-dependent, improved isolated word recognition for slow, loud, and Lombard effect when compared to a neutral trained system.

When the training corpus consists of *only* neutral speech, we propose generating synthetic stressed speech for training. The synthetic stressed speech is obtained here by perturbing the available neutral speech data using hidden Markov model (HMM) perturbation models which have been trained with the wide range of natural variations that exist between neutral and stressed speech parameters. The work in this paper differs from the approaches mentioned previously in that: (1) it eliminates the need for collecting stressed tokens for a particular speaker for training, and instead employs synthetically generated stressed speech for training, (2) it applies the knowledge of how other speakers modify their speech under stress to the neutral speech of new input speakers, and (3) a much larger number of training tokens can be made available. The latter is due to the HMM regenerative property which can be used to produce an unlimited number of perturbation vectors to modify neutral speech, and hence generate stressed synthetic speech.

In this paper, HMMs are employed for two purposes. It is important to clearly distinguish between the two applications. In the first application, HMMs are used for modeling speech parameter variations under stressful conditions, and then for regenerating observations which are statistically equivalent to the training data. These observations are employed for stressed speech synthesis. In the second application, the HMMs are employed for isolated word recognition of neutral and stressed speech.

The remainder of this paper is organized as follows. Section 2 presents the modeling and HMM training of speech parameter variations between neutral and stressed speech. In Section 3, we present the framework for speaking style modification of input neutral speech using the trained HMM perturbation models. Section 4 discusses the HMM topology, and the feature set used for training the recognizer. In Section 5, the results of employing synthetic stressed speech for training are presented. Finally, conclusions are drawn in Section 6.

2. STRESS PERTURBATION MODELS

This work is based on a previously proposed approach for modeling variations in speech parameters under stress using hidden Markov models (HMMs) [4]. The variations in pitch contour, voiced duration, and spectral contour were modeled for the purpose of stressed speech synthesis. Here, the generated synthetic stressed speech is employed for improved recognition performance. In the training

phase, the HMM models were trained with the variations that occur between neutral and stressed speech rather than with actual parameter values since this will vary from speaker-to-speaker. The stressed speaking styles evaluated were angry, loud, and Lombard effect. The following five separate perturbation models were trained for each stressed condition : (i) voiced duration variation, (ii) pitch contour perturbation, (iii) derivative of pitch contour perturbation, (iv) explicit state occupancy for pitch-perturbation HMM, and (v) spectral contour mismatch.

The advantages of our modeling approach are the following: (i) variations in actual speech parameters are being modeled, and therefore the models are not specific to a text-to-speech system, or to a voice coder, (ii) the approach is not dependent on a particular speaker, phoneme class, or word, (iii) the HMM models represent the wide range of parameter variations that exist between neutral and stressed speech (not a fixed perturbation vector), (iv) the HMM models can reproduce unlimited observation sequences with the same statistical properties as the training data (due to the regenerative property of HMMs), and hence a single neutral word can be perturbed in more than one way, and finally, (v) a larger database of stressed synthetic speech can be generated from an originally smaller neutral data set.

After training, and at the perturbation stage, the HMM perturbation models are employed to statistically generate perturbation vectors possessing the same statistical properties as the training data which are used to modify the speaking style of isolated neutral words. Due to the regenerative property of the HMM, more than one perturbation vector can be generated for each neutral word, resulting in stressed synthetic speech with different levels of stress. Hence, the HMM modeling approach allows for a broader representation of the variations under stress than a fixed feature transformation approach which was proposed in an earlier work [5].

3. GENERATING SYNTHETIC STRESSED SPEECH FOR TRAINING

In this approach, we produce synthetic speech tokens for use in training. The HMM-based perturbation models are integrated into a single overall algorithm employing pitch[1], duration, and spectral contour perturbation, as shown in Figure 1, in order to generate stressed speech from neutral speech. The voiced duration distribution, pitch perturbation derivative, and state occupancy model are combined into a single algorithm to generate pitch perturbation profiles. In order to generate a pitch perturbation profile, three steps are necessary: first, the total number of observations to be produced by the whole HMM sequence is determined, second, the number of observations to be generated by each HMM state is determined, and finally, a procedure to order these observations is established. After the duration and pitch perturbation

[1] Although current HMM-based speech recognition systems do not include pitch information as part of the feature set, pitch and prosodic information is important for speech understanding systems.

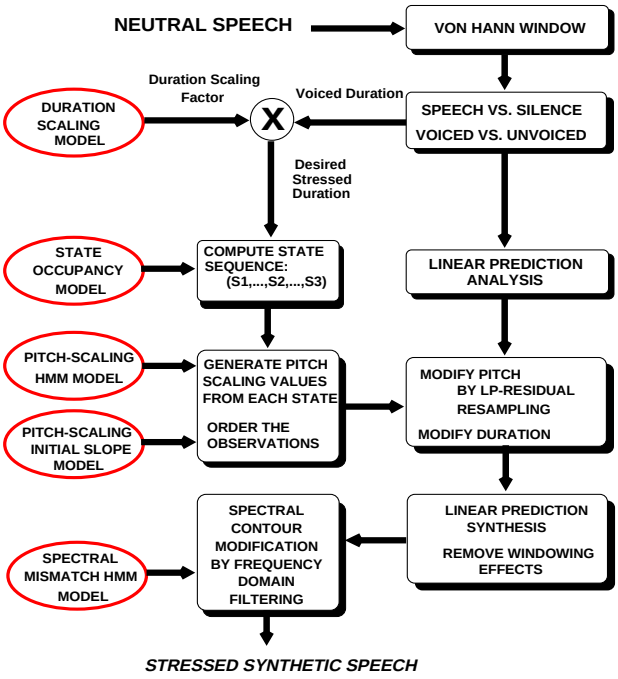

Figure 1: Speaking style modification using HMM-based perturbation models.

step, a spectral perturbation vector is generated for perturbing the spectral contour in the frequency domain. A single spectral perturbation vector is employed across the whole word. A more detailed discussion on each step is presented in [3, 4].

By applying the synthesis steps shown in Figure 1, a total of 6480 synthetic stress tokens are generated from neutral speech for the three stressed speaking conditions (2160 tokens/style). These tokens are generated by perturbing a 24-word vocabulary spoken by a group of nine general American speakers where each word is repeated ten times. The speech database employed in these evaluations is SUSAS : *Speech Under Simulated and Actual Stress* [8]. The vocabulary consists of mono and multi-syllabic words. A common highly confusable vocabulary set of 35 aircraft communication words make up the data base (e.g., /go-oh-no/,/wide-white/,etc). The generated stressed synthetic speech is employed next for training a speech recognition system.

4. RECOGNIZER TRAINING

In this study, two training approaches are considered. The first approach assumes that no neutral or stressed training speech is available from the input speaker, and hence all recognition evaluations are speaker-independent. The second approach assumes that *only* neutral speech is available from the input speaker and that the neutral and synthetic stress models will be adapted to the input speaker. A 24-word HMM-based recognizer is formulated using a variable state number, left-to-right model, with 2 continuous mixtures per state. The features used for training and recognition are 8 LPCC, ΔLPCC, energy, and Δenergy.

In the first scheme, the HMM models are trained in a round-robin scheme with eight speakers while the ninth

Testing With	HMM Models Trained with		
	Neutral	Synthetic Stress	Orig. Stress
Neutral	92.13%	x	x
Angry	78.94%	82.00%	85.88%
Loud	82.87%	86.34%	90.51%
Lombard	90.05%	89.81%	92.59%

Table 1: Performance of speaker-independent models trained with neutral (1^{st} column), synthetic stress (2^{nd} column), and original stressed speech(3^{rd} column).

Testing With	HMM Models Trained with	
	Neutral Speech	Synthetic Stress
Neutral	99.31%	x
Angry	80.09%	85.88%
Loud	85.65%	91.67%
Lombard	94.68%	94.91%

Table 2: Performance of speaker-dependent neutral trained models (1^{st} column), and speaker-adaptive synthetic stress trained models (2^{nd} column).

speaker is left for open testing. A total of 10 tokens per speaker are used for training each neutral word model, resulting in 80 training tokens per word for the neutral models. The original stress models are trained with 16 tokens per word, representing all the available data. A total of 80 training tokens per word are employed for training the speaker-independent synthetic stress models.

In the second scheme, the neutral models are speaker-dependent and hence are trained with the neutral speech of all nine speakers, resulting in 90 training tokens per word. The synthetic stressed models are speaker-adaptive and are trained with the perturbed speech of all nine speakers, resulting in 90 training tokens per word.

5. RECOGNITION EVALUATIONS

The recognition evaluations are presented for both speaker-independent and speaker-adaptive models. In each evaluation, the performance of the neutral trained models is compared to stress-dependent trained models. The three stress conditions considered are angry, loud, and Lombard effect. In all evaluations, the models are tested with a total of 1728 tokens, or 432 tokens per style.

5.1 Speaker Independent Recognition

The baseline recognition performance of the speaker-independent neutral trained recognizer is 92.13% when tested with neutral speech. When neutral trained HMMs are tested with angry, loud, and Lombard speech, recognition performance drops to 78.94% for angry, 82.87% for loud, and 90.05% for Lombard effect.

The original stress trained models improve recognition over neutral trained models for all three speaking styles (see 3^{rd} column of Table 1). As expected, training and testing under similar conditions improves recognition performance. However, when original stressed speech is not readily available for training, we propose training with synthetic stressed speech. The synthetic stress trained models outperform neutral trained models for loud and angry speech as shown in Table 1 (2^{nd} column).

5.2 Speaker-Adaptive Recognition

The recognition performance of speaker-dependent neutral trained models of neutral speech is 99.31%. The recognition rates drop by 19.22% for angry, 13.66% for loud, and 4.63% for Lombard. Speaker-adaptive synthetic stress trained models improve recognition for all three speaking styles as shown in Table 2 (2^{nd} column). Detailed results comparing the performance of speaker-dependent neutral trained models to speaker-adaptive synthetic angry trained models when tested with angry speech are illustrated in Figures 2 and 3. A total of 18 tokens per word are employed for recognition. The words are listed in order of confusability (e.g., /six-fix/, /white-wide/). The synthetic stress training method improves recognition of highly confusable words (note the recognition improvements for the words /three-degree/,/fix-six/,/go-oh-hello-zero/).

5.3 Discussion

This approach generates synthetic stressed speech for training by modifying pitch contour, duration, and spectral contour of neutral speech. Perturbing actual speech parameters as opposed to cepstral parameters, for example, results in a more general approach that is applicable to any speech recognition or understanding system, not just MFCC based systems.

By comparing the performance of both speaker-dependent and speaker-adaptive approaches, we conclude that :

- Speaker-dependent neutral trained models achieve better recognition rates than speaker-independent neutral trained models for all four conditions tested (neutral, angry, loud, and Lombard condition). This indicates that speech produced under stressed conditions possess speaker specific traits.
- High error rates result when testing neutral trained models with stressed speech, indicating that stressed speech possess additional characteristics which are absent in neutral speech.
- Synthetic stress trained models achieve higher recognition rates than neutral trained models especially for highly confusable words.
- Adapting the synthetic stress trained models to the input speaker achieves the highest recognition rates. Therefore, we conclude that under stressed conditions, speakers possess certain similar traits, and hence the knowledge of how a group of speakers modify their speech characteristics under stress can be employed to improve speaker-independent stressed speech recognition performance.
- Speaker-adaptive models (2^{nd} column of Table 2) outperform original stress trained models (3^{rd} column of Table 1). Recall that the original stress models are trained with actual stressed speech from a group of speakers (excluding the test speaker). The

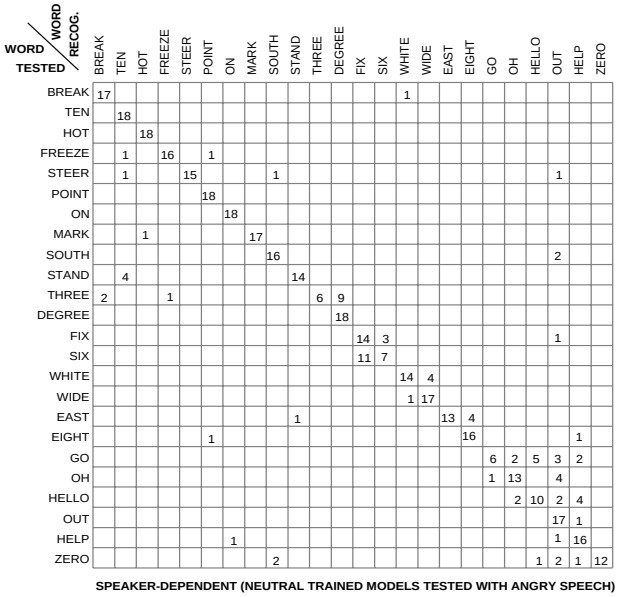

Figure 2: Sample confusion matrix for speaker-dependent neutral trained models tested with angry speech.

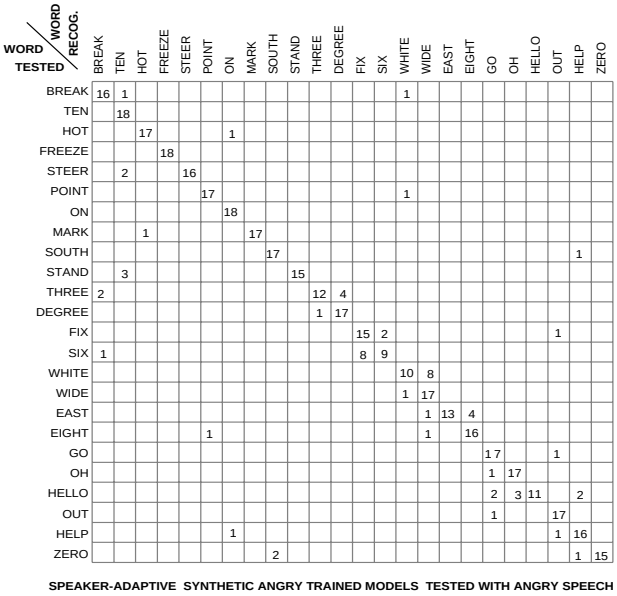

Figure 3: Sample confusion matrix for speaker-adaptive synthetic angry trained models tested with angry speech.

adaptive models, however, are trained with synthetic stressed speech generated by applying the statistical variations that exist under stress onto neutral speech (including that of the test speaker). Therefore, modeling the parameter deviations that exist under stress and employing that knowledge to adapt the recognition models to a new speaker achieves a better performance than directly training the recognizer with the actual stressed data. Our training method clearly eliminates the need for collecting stress training data from the input speaker.

6. CONCLUSIONS

This paper has presented the first approach for improving stressed speech recognition *by generating synthetic stressed speech for training.* Two different training approaches were considered. The first approach consisted of speaker independent training and recognition, and the second approach was speaker-adaptive. Both training methods improve the recognition of angry, loud, and Lombard effect speech. Speaker-dependent neutral trained models outperformed speaker-independent neutral trained models when tested with either neutral or stressed speech. Speaker-adaptive synthetic stress trained models outperformed all other models including the speaker-independent original stressed trained models for all three stressed conditions.

References

[1] S. E. Bou-Ghazale. *Analysis, Modeling, and Perturbation of Speech Under Stress With Applications to Speech Synthesis and Recognition.* PhD thesis, Duke University, Durham, N.C., Nov. 1996.

[2] S. E. Bou-Ghazale and J. H. L. Hansen. Duration and spectral based stress token generation for HMM speech recognition under stress. In *Proc. IEEE-ICASSP*, pp. 413–416, Adelaide, South Australia, Apr. 1994.

[3] S. E. Bou-Ghazale and J. H. L. Hansen. HMM-based stressed speech modeling with application to improved synthesis and recognition of isolated speech under stress. *IEEE Trans. on Speech and Audio*, Dec. 1996. In Review.

[4] S. E. Bou-Ghazale and J. H. L. Hansen. Synthesis of stressed speech from isolated neutral speech using HMM-based models. In *ICSLP*, pp. 1860–1863, Philadelphia, Pennsylvania, Oct. 1996.

[5] S. E. Bou-Ghazale and J.H.L. Hansen. Generating stressed speech from neutral speech using a modified CELP vocoder. *Speech Communication*, 20:93–110, Nov. 1996. Special Issue: "Speech under Stress".

[6] Y. Chen. Cepstral domain stress compensation for robust speech recognition. In *Proc. IEEE-ICASSP*, pp. 717–720, Dallas, Texas, Apr. 1987.

[7] J. H. L. Hansen. Morphological constrained enhancement with adaptive cepstral compensation (MCE-ACE) for speech recognition in noise and Lombard effect. *IEEE Trans. on Speech and Audio*, 2:598–614, Oct. 1994.

[8] J. H. L. Hansen and S. E. Bou-Ghazale. Getting started with SUSAS: A speech under simulated and actual stress database. In *EUROSPEECH*, Rhodes, Greece, Sept. 1997.

[9] J. H. L. Hansen and O. N. Bria. Lombard effect compensation for robust automatic speech recognition in noise. In *ICSLP*, pp. 1125–1128, Kobe, Japan, Nov. 1990.

[10] J. H. L. Hansen and M. A. Clements. Stress compensation and noise reduction algorithms for robust speech recognition. In *Proc. IEEE Inter. Conf. on Acoustics, Speech, and Signal Processing*, pp. 266–269, Glasgow, Scotland, May 1989.

[11] R. P. Lippmann, E. A. Martin, and D. B. Paul. Multi-style training for robust isolated-word speech recognition. In *Proc. IEEE-ICASSP*, pp. 705–708, Dallas, Texas, Apr. 1987.

[12] B. D. Womack and J. H. L. Hansen. Classification of speech under stress using target driven features. *Speech Communication*, 20:131–50, Nov. 1996. Special Issue : "Speech under Stress".