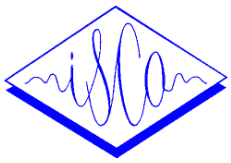

ESCA-NATO Workshop on
Speech under Stress
Lisbon, Portugal
September 14-15, 1995

Improving Recognition and Synthesis of Stressed Speech via Feature Perturbation in a Source Generator Framework

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

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

ABSTRACT

The objectives of this work are two fold, consisting of improving both speech recognition and synthesis of speech under stress. Improved recognition is achieved by generating simulated-stress tokens to replace neutral data used in the recognizer training phase. The second goal is directed at generating stressed speech **from** neutral speech. This is accomplished by formulating speech parameter models for *angry, Lombard effect.* and *loud* speaking conditions, and perturbing the parameters of neutral speech. The studies/evaluations conducted are based on a previously established stress database, called *SUSAS (Speech Under Simulated and Actual Stress).* Results show that the token generation training method improved isolated word recognition by an overall average of 15% when compared to neutral trained isolated word recognition [2]. Results from formal listener evaluations of stress perturbed neutral speech show successful classification rates of 87% for angry speech, 75% for Lombard effect speech, and 92% for loud speech.

1. INTRODUCTION

Several factors degrade the performance of speech recognition systems as well as impair the quality and intelligibility of voice communications. These include (1) speech spoken in noise (i.e., Lombard effect [9]), (2) task-workload (flying an aircraft, operating an automobile), (3) variable speaking style (e.g., speech spoken loud, soft, fast, slow, etc.), and (4) inter-speaker differences. When a speaker is exposed to stress, or wishes to communicate emotional state, production variations will occur in some regular manner. The physical variations are reflected in the speech parameters, such as pitch, spectral slope, and formant structure. If models could be developed to characterize the production system under stress, then such models could be integrated within speech recognition and speech synthesis systems to improve recognition of speech under stress, and the naturalness of synthetic speech.

This research is aimed at identifying indicators of stress for synthesis and recognition, and formulating mathematical/statistical source-generator-based **models** to characterize speech under stress. The source generator framework assumes that speech production can be described as a sequence of speech articulator movements to achieve desired vocal tract target shapes [7]. The collection of speech articulator movements is represented by a sequence of source generators $\gamma_1, ...\gamma_j$, each of which describes an isolated phoneme, a diphone-pair, or some temporal partition. It is suggested that for a neutral word, as shown in Fig. 1, the movement from one source generator to another represents a well defined path in the production space with some degree of natural variations. Under stress conditions, the resulting path is different from that of neutral due to the physical variations that occur when a speaker wishes to communicate emotional state. The physical variations are translated to the speech production feature space which must be modeled.

In this work, the source generator framework will be used to improve stressed speech recognition, and generate stressed synthetic speech from neutral speech data.

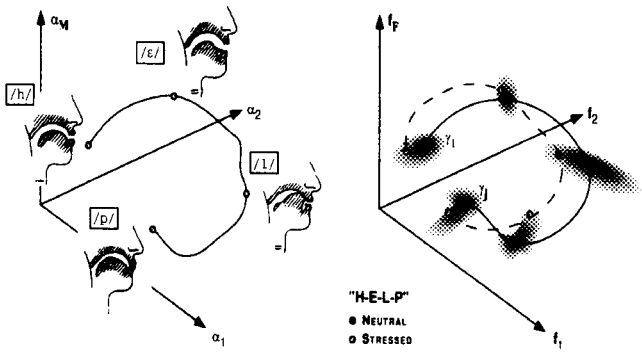

Fig. 1. Neutral and stressed source generator paths in the speech production space.

2. TOKEN GENERATION FOR SPEECH RECOGNITION UNDER STRESS

The performance of a speech recognition system degrades if the recognizer is not trained and tested under similar speaking conditions [4], [5], [6], [8]. It is desired to formulate a procedure for artificially generating stressed speech tokens for hidden Markov model (HMM) training to achieve improved automatic recognition of speech spoken under stressful conditions. The motive for generating these 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. 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.

For this study, statistical models are first developed for duration and Mel-frequency cepstral coefficients (MFCC) for all source generators under each stressed speaking style (e.g., slow, loud, and Lombard). The duration of a source generator is assumed Gaussianly distributed, and is characterized by its mean and variance. The MFCCs of a source generator are characterized by a mean vector calculated across all tokens. When an input neutral word is presented to the algorithm for training, it is desired to perturb the extracted neutral parameters so that they possess stressed speech characteristics. The duration is adapted by varying the number of extracted frames from the speech signal. This is accomplished by maintaining a fixed-length Hamming analysis window while varying the skip rate between successive analysis frames. The MFCC are perturbed according to the following relation

$$C_l^{(s^*)}(j) = [\rho_{mean}(j)_l \times C_l^{(s=Ntr)}(j)],$$

where the mapping factor, $\rho_{mean}(j)_l$, is given by,

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

Here, $C_l^{(s=Ntr)}(j)$ represents the original neutral MFCC extracted for the j^{th} source generator of the input word, and l spans the number of extracted MFCC per frame. The perturbed MFCC is denoted by $C_l^{(s=Ntr)}(j)$. While the perturbation factor, $\rho_{mean}(j)_l$, is represented by the ratio of neutral and stressed mean MFCC. The perturbed parameters are then used for training a 5-state left-to-right hidden Markov model recognizer.

A discrete observation 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. 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 token generation training method improved 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.

3. STRESSED SPEECH SYNTHESIS WITH APPLICATION TO CELP

Next, we wish to consider the prospect of similar feature perturbation for imparting stress on neutral speech. It is desired to apply the source generator based approach to a previously formulated code-excited LP (CELP) speech coder as developed in [1], [3]. Using CELP will allow us to evaluate the ability of a speech coder to properly model and transmit the stressed content present in a speech utterance. By employing CELP however, the features and perturbations will be limited to those available within the CELP framework.

In our application, two assumptions are made: (1) the text of the input word is known to an input parsing algorithm, and (2) the system has access to the CELP synthesizer. The first assumption ensures that all utterances of a word are parsed into the same sequence of source generators. The second assumption is required in order to have access to the transmitted parameters prior to synthesis.

3.1 Neutral and Stressed Speech Model Formulation

CELP perturbation models were developed for line-spectral-pair (LSP) parameters, pitch delay, and pitch gain. LSP models are generated for all source generator classes, while pitch delay models are generated for voiced source generators only. The LSPs of a source generator are assumed to be Gaussianly distributed and are characterized by their sample mean and variance. Pitch delay profiles are modeled by a third order polynomial, and overall mean duration. The gain is characterized by a scalar representing the overall average gain. Perturbation models are formulated for each speaking style (e.g., *neutral*, *angry*, *Lombard effect*, and *loud*). These models will be used for perturbing speech parameters of neutral input speech utterances, resulting in stressed synthetic speech.

3.2 Perturbation of Neutral Speech Parameters

Perturbation of input neutral speech can be achieved as follows. Given neutral input speech, speech parameters are extracted through the CELP analyzer. The input speech is also processed by the parsing procedure to identify the source generator classes and boundaries. The parser is a hidden Markov model (HMM) based phoneme classifier. Once the source generators, $\gamma_1, \gamma_2, ..., \gamma_j$, are identified, the perturbation algorithm selects the proper neutral and stressed source generator based models. The extracted neutral speech parameters, the source generator boundaries, and the LSP, pitch delay and pitch gain models, are presented to the perturbation algorithm where the actual mapping is implemented. The perturbation can be done in one of four ways: (1) Line-spectral pair Mapping (LM), (2) Pitch Delay Mapping (PDM), (3) Gain Mapping (GM), or (4) Line-spectral pair, Pitch delay, and Gain Mapping (LPGM). A linear mapping is applied to the LSPs, a non-linear pitch profile mapping is applied to the pitch delay, and a scalar translation is applied to the pitch gain. Under stressed speech conditions, the LSPs are assumed to vary linearly according to the following linear transformation

$$\vec{L}_{\gamma_i(S)} = \vec{a}_{\gamma_i} \odot \vec{L}_{\gamma_i(N)} + \vec{b}_{\gamma_i}$$

where $\vec{L}_{\gamma_i(N)}$ represents the LSP distribution of the source generator γ_i under neutral conditions, $\vec{L}_{\gamma_i(S)}$ represents the LSP distribution of γ_i under stress conditions, and $\odot$ is a point-by-point multiplication. The linear mapping coefficients $\vec{a}_{\gamma_i}$, and $\vec{b}_{\gamma_i}$ are unique for each source generator class under each stress style. Pitch delay perturbation is obtained by the following non-linear mapping

$$\overrightarrow{PDI}_{(p)} = \overrightarrow{PDI}_{(N)} \oslash (\overrightarrow{PDM}_{(S)} \oslash \overrightarrow{PDM}_{(N)}),$$

where $\overrightarrow{PDI}_{(N)}$ is the input *neutral* pitch delay, $\overrightarrow{PDM}_{(S)}$ and $\overrightarrow{PDM}_{(N)}$ are the *stressed* and *neutral* pitch delay models after duration adjustment, and $\overrightarrow{PDI}_{(p)}$ is the resulting *perturbed* pitch delay input. The $\oslash$ and $\odot$ represent a point-by-point multiplication and division. The perturbed parameters are used for synthesis, and the resulting speech is the original input word spoken under stress conditions.

3.3 Perturbation Algorithm Results

The evaluations of the proposed modeling scheme consisted of parametric analysis, and a set of formal pairwise subjective listening tests. Formal subjective listening results of CELP coded stressed speech demonstrated a slight loss in CELP's ability to reproduce stressed speech (Table I). Certain speaking styles were better reproduced by CELP than others. CELP coded angry speech, for example, was judged 8% of the time as sounding less angry than the original angry speech. While CELP coded loud speech was judged only 3% of the time as sounding less loud than its original token. Finally, CELP coded Lombard speech was better reproduced than any other stressed style with only a 2% loss in CELP's ability to reproduce it.

Detailed listener results of the generated loud speech are presented in Fig. 2. The results show that the original loud speech was classified 85% of the time as sounding loud, and the original neutral speech was classified 5% of the time as sounding loud. The results also show that LSP modification alone (LM) does not convey the emotional state of speech to the listener. Gain modification (GM) alone also does not convey sufficient stress cues. In addition, the results show that PDM perturbed loud speech was classified 62% of the time as sounding loud, while the LPGM perturbed loud speech was classified 92% of the times as sounding loud. The two most effective algorithms were pitch delay mapping (PDM) and the combination of LSP, pitch delay, and pitch gain modification (LPGM). Note that the LPGM perturbed loud speech resulted in even a higher classification level than the original loud speech (7% higher).

A summary of the overall results for *angry*, *Lombard effect*, and *loud* speaking styles is given in Fig. 3 for PDM and LPGM perturbed speech. The LPGM perturbed loud speech resulted in the highest classification rate, and was perceived 92% of the time as sounding loud. LPGM perturbed angry was perceived 87% of the time as sounding angry. Finally, LPGM perturbed Lombard was classified 75% of the time as appearing to have Lombard effect characteristics. This suggests that listeners employ a combination of excitation and vocal tract acoustic cues in their perception of stress/emotional speech content.

The *LSP*, *pitch*, and *gain* (LPGM) modified speech was better able to convey the emotional state of speech to the listener than *pitch delay* (PDM) perturbation alone. In addition, listeners noted that LPGM modified speech had a better quality than PDM modified speech. Listeners also indicated a noticeable but unobjectionable loss in quality in the generated LPGM modified stressed speech when compared to original neutral speech.

4. DISCUSSION AND CONCLUSION

This study has considered feature perturbation within a source generator framework for improving

Input Speech	Loss in Performance
Angry	8%
Lombard	3%
Loud	2%

TABLE I
Loss in CELP's ability to reproduce stressed speech.

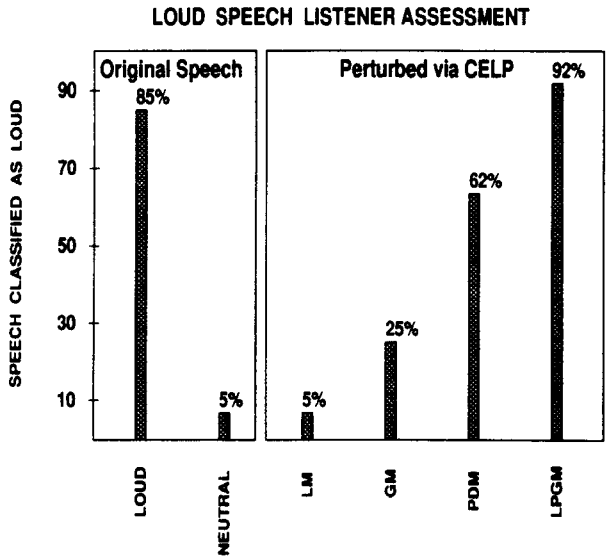

Fig. 2. A detailed bar graph from a subjective listening test illustrating (1) the percent of times the original *neutral* and *loud* speech were classified as sounding loud, and (2) the percent of times the generated perturbed speech was chosen as more loud then the original neutral.

stressed speech recognition, and generating stressed speech from neutral input data. A new approach for generating simulated stressed training tokens has been presented and demonstrated for a discrete observation hidden Markov model recognizer in an isolated word scenario. The proposed training method improved stressed speech recognition by an overall average of 15%. Second, a new approach for producing speech which possess stressed speech features has been presented and demonstrated using a 4800 bps CELP vocoder. Results showed that formant locations alone or overall gain alone were not sufficient relayers of stress; while pitch delay was a more successful indicator of the emotional state of the speaker. The combination of LSP, pitch delay, and gain modification (LPGM) produced the highest subjective listening results. The results suggest that feature perturbation in a source generator framework can be successful in improving the overall performance of recognition and synthesis of stressed speech.

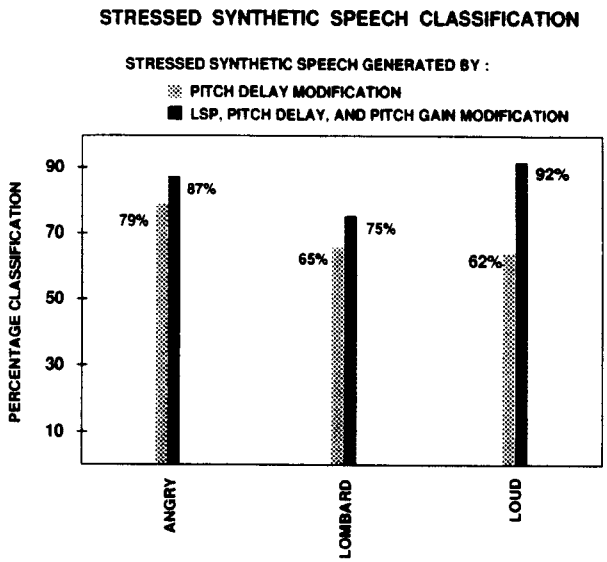

Fig. 3. A summary of the subjective listening test illustrating the performance of two stress perturbation algorithms : PDM, and LPGM. The bar graph shows the percent of times the generated perturbed speech is correctly classified.

5. REFERENCES

[1] B.S. Atal and M.R. Schroeder. Stochastic coding of speech signals at very low bit rates. In *Proc. IEEE Int. Conf. Communications*, page 48.1, May 1984.

[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 Int. Conf. on Acoustics, Speech and Signal Processing*, pages 413–416, Adelaide, South Australia, April 1994.

[3] J. P. Campbell, V. C. Welch, and T. E. Tremain. An expandable error-protected 4800 bps celp coder (u.s. federal standard 4800 bps voice coder). In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 735–738, Glasgow, Scotland, May 1989.

[4] 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.

[5] 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.

[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. Adaptive source generator compensation and enhancement for speech recognition in noisy stressful environments. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 95–98, April 1993.

[8] R. P. Lippmann, E. A. Martin, and D. B. Paul. Multistyle 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.