A Source Generator based Modeling Framework for Synthesis of Speech Under Stress

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 objective of this paper is to formulate an algorithm to generate stressed synthetic speech **from** neutral speech using a source generator framework previously employed for stressed speech recognition [2, 3]. The following goals are addressed (i) identify the most visible indicators of stress as perceived by the listener in stressed speaking styles such as loud, Lombard effect and angry, (ii) develop a mathematical model for representing speech production under stressed conditions, and (iii) employ the above model to produce emotional/stressed synthetic speech **from** neutral speech. The stress modeling scheme is applied to an existing low-bit rate CELP speech coder in order to investigate (i) the coder's ability and limitations reproducing stressed synthetic speech, and (ii) our ability to perturb coded neutral speech parameters at the synthesis stage so that resulting speech is perceived as being under stress. Two stress perturbation algorithms are proposed and evaluated. Results from formal listener evaluations show that 87% of neutral perturbed speech was indeed perceived as stressed.

1. INTRODUCTION

In the field of speech modeling, there have been numerous studies directed at high quality speech coding and methods for natural speech synthesis. In the area of coding, the majority of this work has focused on maintaining the level of input quality, speaker identity traits, and message intelligibility. There has been little work however, directed at understanding what impact vocoders have on the stress or emotional content of a speech utterance. In some voice telephone applications, the ability to accurately perceive stress or emotion can be critical. For example, in message sorting of 911 emergency telephone traffic, or monitoring of air traffic control communications, the ability of a voice coding method to clearly represent speaker stress or emotional state could result in more effective emergency response measures. Improvements in voice coding methods for such applications would also suggest areas for advancement in general speech synthesis, contributing to an increased overall level of naturalness.

In this study, the problem of stressed speech synthesis from neutral input speech data is addressed. The problem is considered from a general speech modeling perspective using a source generator framework previously considered for robust speech recognition [3, 6]. In Sec. 2, we discuss the source generator model framework (see Fig. 1), and illustrate how it can be employed for synthesis of speech under stress. In Sec. 3, the framework is employed within an existing low-bit rate speech coder for synthesis of stressed speech. Next, evaluations are presented which identify limitations of the coding method to reproduce stressed synthetic speech, and our ability to perturb neutral features so that the resulting speech will be perceived as being under stress. Finally, in Sec. 5, we summarize and draw conclusions from our study.

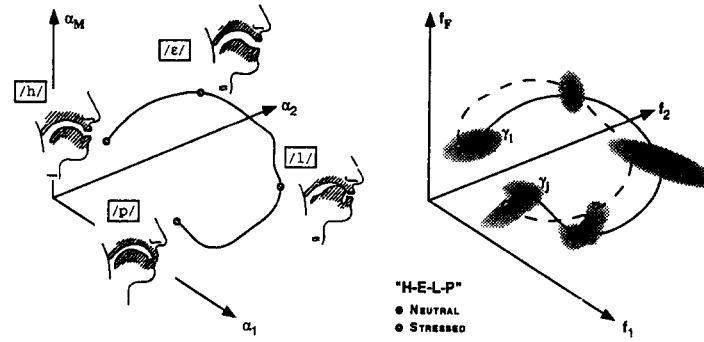

Figure 1: Neutral and stressed source generator paths in the speech production space.

2. SOURCE GENERATOR FRAMEWORK

The proposed framework assumes that speech production can be described as a sequence of speech articulator movements to achieve desired vocal tract target shapes. The collection of speech articulator movements is represented by a sequence of source generators $\gamma_1, \ldots \gamma_j$, each of which describes an isolated phoneme, a diphone-pair, or some temporal partition (Fig. 1).

The source generator approach was previously employed [3] in a framework for generating simulated stressed tokens which were used for training a robust stressed speech recognizer. In this approach, both duration and mel-cepstral parameters were altered to statistically resemble a stressed speech token. The token generation training method improved isolated word recognition by an overall average of 15% for slow, loud and Lombard effect when compared to neutral trained isolated word recognition.

It is desired to employ a similar approach to perturb neutral speech parameters so that the resulting synthetic speech is perceived as being under stress. A modified framework is necessary for the application of synthesis, since the speech production relayers which contribute most to reduced recognition performance may not be those which are most visible in synthesis to the listener.

It is suggested that for a neutral word, the movement from one source generator to another represents a well defined path

0-7803-2431-5/95 $4.00 © 1995 IEEE

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 as shown in Fig. 1.

In the most general case, it is assumed that speech production under a given stress condition can be represented by a translation of the speech vector $\vec{S}$, through a transformation of the speech production feature set $\vec{\Phi}$. Let us assume that the stress condition is speech spoken under angry conditions. Then the following transformation is suggested

$$g(\vec{S}_{Ntrl}) \rightarrow \vec{S}_{Angry}$$

$$G(\vec{\Phi}_{Ntrl} : \vec{\Phi}_{exc}, \vec{\Phi}_{vt}) \rightarrow (\vec{\Phi}_{Angry} : \vec{\Phi}_{exc}, \vec{\Phi}_{vt})$$

where the translation $g(\cdot)$ operates on the speech signal, while the transformation $G(\cdot)$ operates on the speech production feature set $\vec{\Phi}_{Ntrl}$. Here, $\vec{\Phi}_{exc}$ represents potential speech excitation features due to glottal source (e.g. mean pitch, pitch range, pitch contour, glottal waveform spectral slope, glottal source spectrum/amplitude), unvoiced excitation (e.g. frication), and mixed excitation features (e.g. due to plosives). $\vec{\Phi}_{vt}$ represents vocal-tract features such as formant locations, bandwidths, formant transitions, etc.

The proposed transformations used for imparting *stress* on *neutral* speech could perturb global features at the word or phrase level (e.g., average pitch), or at the phonemic level (e.g., instantaneous pitch). This general production feature space is considered within the context of a CELP vocoder.

3. APPLICATION OF STRESS PERTURBATION via CELP CODING

Next, we consider the ability to impart stress/emotion on neutral speech within a CELP vocoder [1, 4]. It is therefore necessary to consider (i) speech modeling in CELP, in order to identify potential parameters for stress perturbation, and (ii) the effect of stress on CELP modeling parameters to determine which features are most effective as stress relayers.

3.1 Speech Modeling in CELP

The 4.8 kbps code-excited linear predictive vocoder consists of an analysis and synthesis sections. Analysis consists of three basic functions: (i) short-term linear prediction (related to spectral envelope), (ii) long-term adaptive codebook search (related to pitch period of voiced speech), and (iii) residual "codebook" search. The synthesis component consists of the same three functions (performed in reverse order) with the addition of a postfilter for output speech enhancement. The CELP algorithm processes input speech on a frame-by-frame basis, with parameters extracted and quantized prior to transmission across the channel. The transmitted parameters are the *stochastic* codebook index and gain, the *adaptive* codebook index and gain, and 10 line spectral parameters (LSP). At the synthesis stage, the received parameters are decoded and employed in the synthesis model.

3.2 CELP Perturbation Models

An extensive evaluation was performed on CELP's parameters to establish perturbation models for speech under stress.

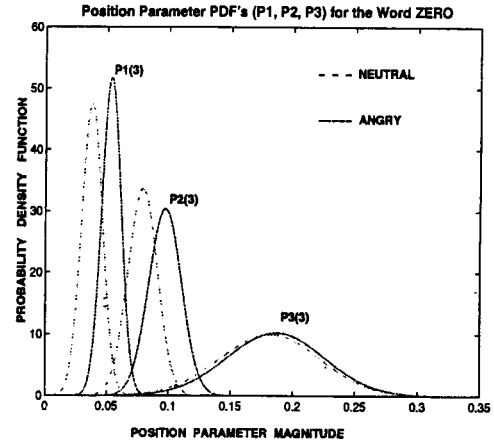

Figure 2: Distribution of the first 3 position parameters for the /e/ sound (i.e., source generator γ_3) of the word ZERO under neutral and angry speaking conditions.

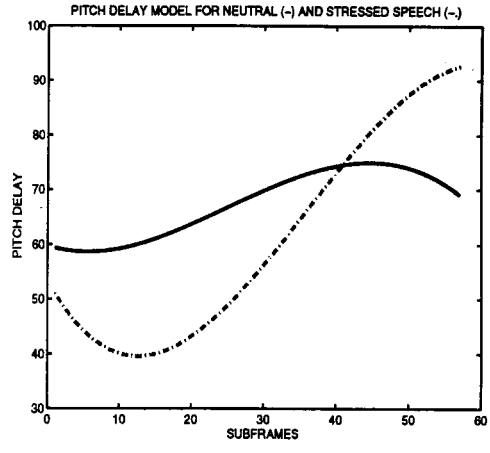

Figure 3: Pitch delay MODELS for the word *go* under original neutral conditions (solid) and original angry conditions (dashed).

From these studies, it was determined that the most effective relayers of stress are the LSPs (updated on a frame basis), and the adaptive long delay pitch parameters. The LSP perturbation model is characterized here by a conditional Gaussian distribution (i.e., always ≥ 0); while pitch delay is modeled by a 3^{rd} order polynomial using a least-squares procedure. The formulated LSP, pitch, and gain models are unique for each source generator and for each speaking style. Sample LSP distributions (position parameters) are shown in Fig. 2 across all tokens of the word *zero* under neutral and angry conditions. This figure demonstrates the statistically significant shift in the 1^{st} and 2^{nd} position parameters of the third source generator (γ_3) under angry conditions. Pitch delay profiles were also shown to vary significantly across stressed speaking conditions. Fig. 3 illustrates the clear shift between neutral and angry pitch delay profile *models*.

3.3 Algorithm Framework

The approach for perturbing neutral parameters to produce stressed synthetic speech within the CELP coder is to transform (i) the vocal-tract features, represented by the LSPs, and

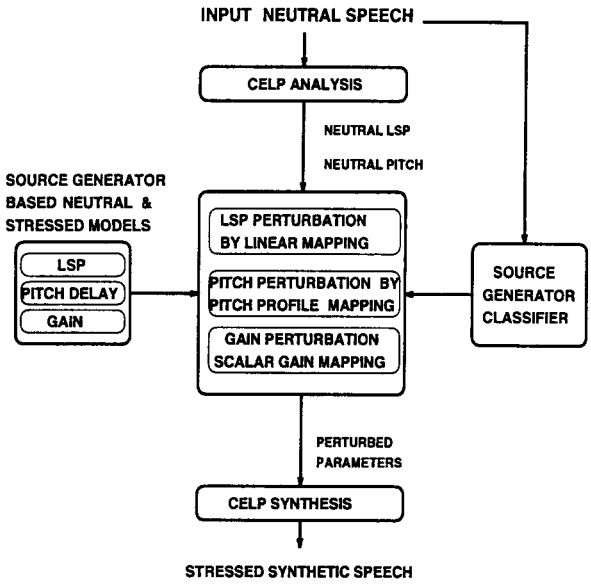

Figure 4: Flow graph for generating stressed synthetic speech via the CELP coder.

(ii) the excitation features, represented by the adaptive pitch delay and pitch gain parameters.

In order to implement the speech perturbation algorithm, source generator-based models for LSPs and pitch parameters are developed for each speaking style. Recalling that the neutral (L_N) and stressed (L_S) LSP distributions are assumed Gaussian, the LSP mapping from neutral to stressed speech is given by the linear transformation $L_S = aL_N + b$. The linear projected coefficients will be unique for each source generator under each speaking style.

Pitch profile models are generated for voiced source generators. Pitch delay perturbation is a point-by-point projection given by $\vec{P}[n] = \vec{I}[n] \cdot \times (P\vec{D}M_S \cdot / P\vec{D}M_N)$, where $\vec{I}$ is the input pitch profile, $P\vec{D}M_S$ is the stressed pitch delay model, $P\vec{D}M_N$) is the neutral pitch delay model, and $\vec{P}$ is the resulting perturbed pitch profile. The $\cdot\times$ and $\cdot/$ represent a point-by-point multiplication and division. Pitch gain model is represented by the overall average gain across a source generator class.

Once modeling is completed, perturbation of speech parameters can be done as shown in Fig. 4. In step one (i) we process neutral speech through the CELP analyzer, using a 180 point analysis window (45 samples/subframe), with quantization disabled, and (ii) parse the input speech through an HMM-based phoneme classifier to label source generator boundaries. Given detected source generator sequence, the perturbation algorithm selects the proper stress generator models. Then, at the CELP synthesis stage, prior to actual synthesis, we perturb (i) the LSP structure using the linear mapping, (ii) the pitch delay profile using the pitch profile mapping, and (iii) pitch gain by a multiplicative scalar mapping.

4. EVALUATIONS

To evaluate stressed speech synthesis using a CELP coding framework, a formal listening evaluation was conducted on the perturbed stressed speech, and the resulting CELP features were analyzed with respect to actual stressed features. The evaluations include analysis of line spectral pairs, pitch delay and pitch gain, as well as effects of quantization and analysis frame size on stressed speech coding.

Input Speech IS	Speech Judged As	*% of times*
Original Angry	Angry	96.97%
CELP Angry	Angry	88.30%

Table 1: Ability of the CELP coder to reproduce stressed speech.

4.1 Subjective Listening Test

It is clear that subjective assessment of speech under neutral and stressed conditions is important. Our criterion for this assessment is based on subjective listener evaluations. The main goals are to determine (i) the ability of listeners to correctly classify speaking style as being neutral or stressed, (ii) the limitations of CELP in reproducing stressed speech, and (iii) the ability of our perturbation algorithm to generate stressed synthetic speech via CELP.

For the first test, the listener heard a sequence of individual words and was prompted to judge whether each word sound neutral or stressed. This evaluation used the original speech, and established a baseline of listener perceived stress.

For the second test, the listener heard a series of word pairs and was asked to pick 1 of 4 choices: (1) first word is more stressed, (2) second word is more stressed, (3) both words are equally under stress, or (4) neither word is under stress. Word pairs consisted of either (a) an original stressed word before and after CELP coding, or (b) an original neutral word and its stressed perturbed CELP coded version. A random word order was used for both listening tests.

4.2 CELP Coding of Neutral and Stressed Speech

From our analysis of CELP features, we note that some of CELP's limitations are its inaccurate reproduction of high frequency formant structure, and its inability to adequately track rapid changes in the speech signal (also confirmed earlier by Kroon [7]). Formal subjective listening results of CELP coded angry speech demonstrated a slight loss in CELP's ability to reproduce stressed speech (see Table 1). When original angry speech was presented to listeners, 97% of the speech was classified as angry. When original angry speech was compared to its CELP coded sequel, 88% of the CELP coded angry speech was judged to be equally angry as the original before CELP coding.

4.3 Perturbation Algorithm Results

Two stress perturbation algorithms were developed and evaluated in this study. The first algorithm generates stressed synthetic speech by pitch delay mapping (PDM) alone; while the second algorithm uses LSP, pitch delay, and pitch gain mapping. Table 2 summarizes listener results for a comparison of the performance of the *LSP*, *pitch*, and *gain* (LPG) modified speech to the PDM modified speech. The LPG perturbation

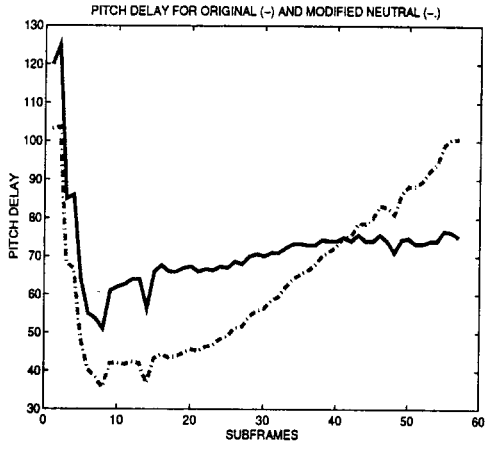

Figure 5: Original neutral pitch delay for the word *go* before (solid) and after (dashed) pitch profile perturbation.

Modified Speech	% judged as more stressed
PDM modified speech	79.0%
LPG modified speech	86.6%

Table 2: Listening test # 2 : Subjective listening test illustrating the percent of times the generated stressed synthetic speech was chosen as more stressed than the original neutral.

was better able to convey the emotional state of speech to the listener than *pitch* perturbation alone. LPG perturbed neutral speech was perceived 87% of the time as sounding stressed. PDM perturbed speech was perceived 79% of the time as stressed. In addition, listeners noted that LPG modified speech had a better quality than PDM modified speech. Listeners also indicated a noticeable but unobjectionable loss in quality in the generated LPG modified stressed speech when compared to original neutral speech.

Fig. 5 shows an example pitch delay profile of a voiced speech segment before and after pitch perturbation. The models used for this mapping were given in Fig.3. It can be seen that the perturbed pitch profile is tracking the angry pitch delay model given in Fig.3.

Fig. 6 shows examples of LSP#1, and LSP#3 plots for the word *break* before and after LSP linear mapping. The difference between the original neutral LSP and the perturbed LSP is manifested near the second source generator region (frames 10 to 22 in Fig. 6). Also, note that the LSP mapping is not identical across the word, but instead unique to the source generator class.

5. DISCUSSION AND CONCLUSION

A new approach for producing synthetic speech which possess stressed speech features has been presented and demonstrated using a 4800 bps CELP vocoder. The conclusions drawn from our studies are: (i) spectrogram analysis revealed that CELP was better able to reproduce the formant structure for *neutral* speech than for *stressed* speech, (ii) mapping either the source excitation, or spectral envelope, produces good quality intelligible speech, but it is the combination of excitation and

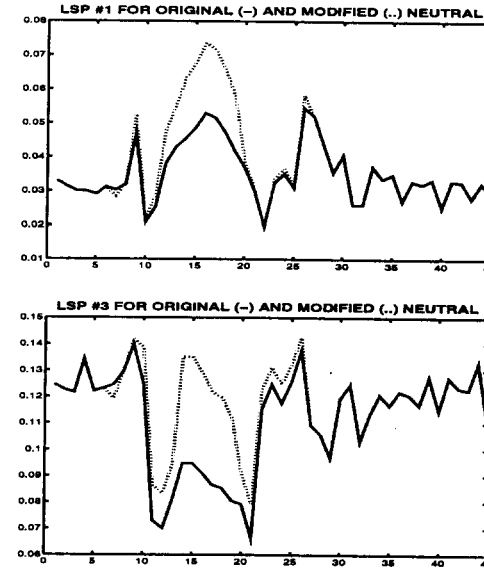

Figure 6: Plots for LSP # 1 (TOP) and LSP # 3 (BOTTOM) of the word *break* under original neutral (solid) and perturbed to resemble angry conditions (dashed).

spectral information transformation that provides the necessary cues to the listener about the speaker's stress state, and (iii) using a higher update rate for the filter's coefficients and a smaller analysis frame improves some of the formant structure at higher frequencies.

The subjective listening tests showed an 8% loss in CELP's ability to reproduce stressed speech. Furthermore, listener's results demonstrated that the perceived stress in the modified speech is noticeable when parameter perturbation is used to modify the speech to sound more stressed (87% of the generated stressed speech was classified as stressed).

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] Sahar E. Bou-Ghazale. " Duration and Spectral Based Stress Token Generation for Keyword Recognition Using Hidden Markov Models ". M.S., Duke University, Dept. of Electrical Engineering, June 1993.

[3] Sahar E. Bou-Ghazale and J.H.L. Hansen. " Duration and Spectral Based Stress Token Generation for HMM Speech Recognition under Stress ", *IEEE ICASSP-94*, pp. 413–416, April 1994.

[4] 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)". *ICASSP - 89*, pages 735–738, 1989.

[5] J.H.L. Hansen, "Analysis and Compensation of Stressed and Noisy Speech With Application to Robust Automatic Recognition," Ph.D. Thesis, Georgia Inst. of Tech., 428 pgs, July 1988.

[6] J.H.L. Hansen, "Adaptive Source Generator Compensation and Enhancement for Speech Recognition in Noisy Stressful Environments," *ICASSP - 93*, vol. II, pp. 95-98, April 1993.

[7] P. Kroon and B.S. Atal. "Strategies for Improving the Performance of CELP Coders at Low Bit Rates". *ICASSP - 88*, pp. 151–154, New York, New York, 1988