SYNTHESIS OF STRESSED SPEECH FROM ISOLATED NEUTRAL SPEECH USING HMM-BASED MODELS

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

Robust Speech Processing Laboratory
Duke University Department of Electrical and Computer Engineering
Box 90291, Durham, North Carolina 27708-0291
http://www.ee.duke.edu/Research/Speech

ABSTRACT

In this study, a novel approach is proposed for modeling speech parameter variations between neutral and stressed conditions and employed in a technique for stressed speech synthesis. The proposed method consists of modeling the variations in pitch contour, voiced speech duration, and average spectral structure using Hidden Markov Models (HMMs). While HMMs have traditionally been used for recognition applications, here they are used to statistically model characteristics needed for generating pitch contour and spectral slope patterns to modify the speaking style of isolated neutral words. An algorithm is developed based on an analysis-synthesis speech model, and HMM pitch and spectral stress characteristics for stress perturbation. Informal listener evaluations of the stress modified speech confirm the HMMs ability to capture the parameter variations under stressed conditions. The proposed HMM models are both speaker and word-independent, but unique to each speaking style. While the modeling scheme is applicable to a variety of stress and emotional speaking styles, the evaluations presented in this study focus on angry, Lombard effect, and loud spoken speech.

1. INTRODUCTION

Modeling speech parameter variations under various stressed speaking conditions is an important problem both for improving the naturalness of speech synthesis, and robustness of stressed speech recognition algorithms. This paper presents a general modeling approach based on HMMs that is speaker and text-independent for representing variations in speech parameters under stressed speaking conditions. Such a model can represent the wide range of natural variations that exist between neutral and stressed speech parameters, and has the ability to regenerate parameters with the same statistical properties as the training data. The regenerated parameters can be used to modify the speaking style of an input neutral word in more than one way, resulting in different levels of stress for the same input word. Hence, the HMM modeling approach allows for a broader representation of the variations under stress than a fixed feature transformation approach which was previously proposed [1].

Traditionally, HMMs have been used for speech recognition. However, in this study, they will be used to model changes in pitch, duration, and spectral tilt that occur under stressed conditions. These models are then used to regenerate parameters which will be used for perturbing neutral speech and hence modify speaking style. The notion of imparting stress or emotion onto input neutral speech is an important research area which is emerging in the field of speech coding/synthesis and text-to-speech synthesis.

While several pitch contour model approaches have been proposed, only two studies have employed HMM-generated pitch contours for speech synthesis [3, 8]. The proposed work here, however, differs from these previous approaches in that: (1) it is less computational, (2) no separate models are devised for high/low falls, high/low rises, fall-rises, and rise-falls; since the whole pitch profile is modeled as a single 3-state HMM, no concatenation of pitch-HMMs is necessary, (3) pitch contour normalization is not required, and (4) our approach makes no explicit use of the phonemic environment. The focus in this study, therefore, is the development of a novel technique for pitch contour, duration and spectral slope generation using HMMs for the purpose of stressed speech synthesis.

2. PARAMETRIC ESTIMATION

The speech parameters studied in this work represent variations that occur from neutral to stressed speech in (i) pitch profile, (ii) average spectral slope, and (iii) voiced speech duration. These parameters are computed as follows. Pitch estimation is based on a modified autocorrelation method as described in [6]. Prior to pitch estimation, the input speech is partitioned into speech/silence, and voiced/unvoiced regions. The speech/silence decision is based on a combination of two measures: peak-to-peak amplitude, and zero-crossing rate. The voiced/unvoiced decision is based on signal-to-residual energy ratio, and normalized autocorrelation. The spectral slope is computed for each speech frame within a word by finding a second order least squares fit to the spectrum. The voiced/unvoiced procedure is used to extract the duration of voiced speech; hence only voiced speech duration is modeled.

The speech data employed for both analysis and synthesis evaluation is a subset from a previously established database called *SUSAS (Speech Under Simulated and Actual Stress)* [4]. Approximately half of the SUSAS database consists of styled data (such as *normal, angry, soft, loud, slow,*

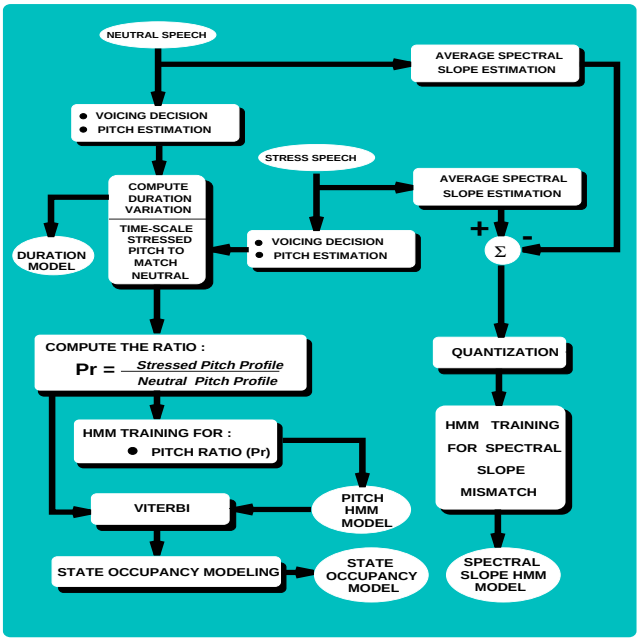

Figure 1: Flow diagram showing duration modeling, HMM training of pitch ratios and spectral slope, and explicit HMM state occupancy modeling.

fast, clear, Lombard effect[1], *etc.*)[2]. A common vocabulary set of 35 aircraft communication words make up over 95% of the data base. Examples include /go-oh-no/, /wide-white/, and /six-fix/. Twelve tokens of each word in the vocabulary were spoken by nine native American speakers for neutral conditions, and two tokens per word for the styled conditions.

3. HMM-BASED SPEECH PARAMETER MODELING

The HMMs serve two purposes in this study. First, they are used to represent the wide range of natural variations that exist in speech production between neutral and stressed speech conditions. Second, due to their regenerative property, they can reproduce unlimited observation sequences with the same statistical properties as the training data. This ability to regenerate a large number of parameter scaling profiles allows a single neutral word to be perturbed in more than one way, which results in different levels or types of stress for the same input word.

3.1 HMM Training

The HMM training data employed consists of 29 utterances. A total of 6264 training vectors were used in training each of the pitch-HMM, spectral-slope-HMM, duration model,

[1] Lombard effect speech was obtained by having speakers listen to 85 dB SPL pink noise through headphones while speaking (i.e., recordings are noise-free).

[2] A portion of this database was donated by Lincoln Labs [7].

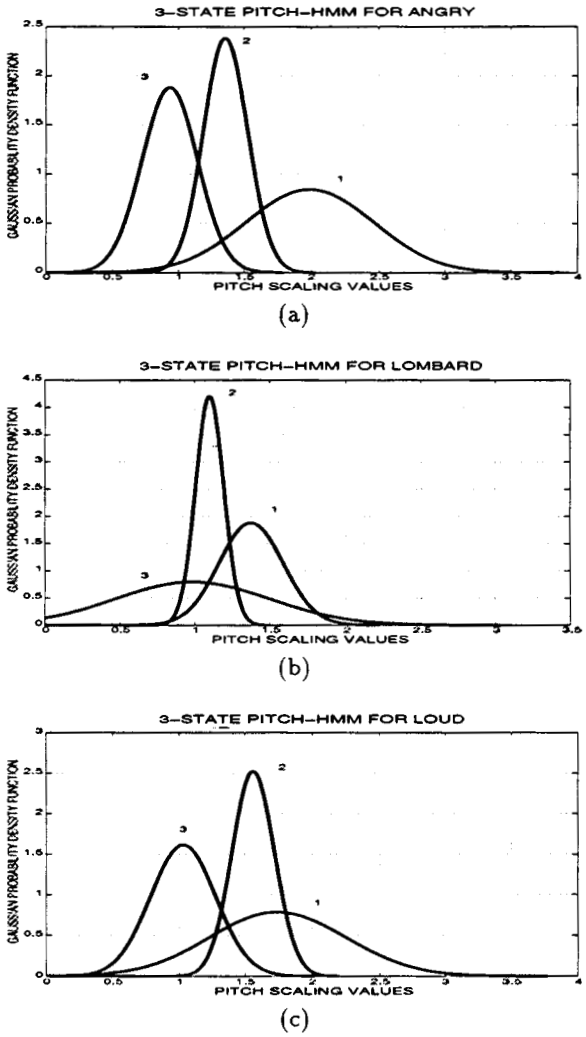

Figure 2: 3-state Pitch-HMM distributions for (a) Angry, (b) Lombard effect, and (c) Loud stress styles.

and state occupancy model for each stress condition. The stressed speaking styles evaluated in this paper are angry, Lombard effect, and loud.

In the training phase, separate HMM models are obtained for pitch contour and average spectral slope. It is also necessary to form separate models for each stressed speaking condition. Voiced duration variation and state occupancy are modeled using probability mass functions (PMF). As shown in the modeling flow diagram of Fig. 1, pitch contour, average spectral slope, and voiced duration are computed simultaneously for an input neutral and stressed word (same speaker, same text). The pitch-HMM model is trained with those pitch variations that occur from the neutral to stressed speaking condition rather than with actual pitch values. This allows pitch-HMM perturbation to be applied to open test-set speakers (i.e., under stressed speaking conditions, it is more important to model the deviation from neutral than the actual min/max/mean since this will vary from speaker-

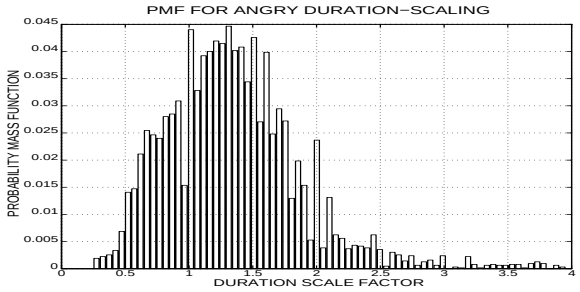

Figure 3: PMF distribution for neutral-to-angry duration transformation.

to-speaker). Pitch variation is represented as a ratio, $\mathbf{P_r}$, of stressed pitch contour to neutral pitch contour. Prior to computing these ratios, the duration of the stressed pitch is scaled to match that of neutral. The pitch profile ratios are then used to train a 3-state HMM model. This scheme allows one to increase or decrease a speaker's pitch by certain factors rather than impose whole pitch contours that are not natural to the input speaker. Gaussianly distributed PDFs of pitch scaling observations corresponding to 3-state pitch-HMMs for angry, loud, and Lombard are shown in Fig. 2. These plots illustrate that pitch varies differently across time depending on the speaking style, and that HMMs are able to model these differences.

Once a pitch-HMM perturbation model is trained with all pitch ratios (6264 training vectors), both the model and training data are submitted to a Viterbi algorithm which computes the optimal state sequence for the observation vectors. This procedure is repeated with every training vector. The resulting state sequences are then used to construct a histogram that represents the duration spent in each state. The state occupancy distribution of a state is conditioned on the time spent in all previous states. Ferguson [2] and others have developed other techniques for state occupancy modeling.

The voiced duration of neutral and stressed words are used for computing the voiced duration model as shown in Fig. 1. Duration variation is modeled as the ratio of stressed-to-neutral voiced duration. These values are then used to construct a duration PMF. An example PMF distribution for neutral-to-angry duration transformation is shown in Fig. 3. Here, the average duration scale factor for neutral-to-angry transformation is approximately 1.373 (i.e., voiced duration is increased 37%).

The average spectral slope of neutral and stressed words, as shown in Fig. 1, are used to compute the average spectral difference. As was mentioned in Sec. 2, the spectral slope is estimated using a 2^{nd} order least squares fit to the spectrum. Instead of training an HMM with vectors containing 513 data points, the 2^{nd} order spectral fit was quantized to 3 points across a 4 kHz bandwidth. These points are sufficient to uniquely describe the 2^{nd} order function, and are used to train a 1-state 3-parameter HMM, in which each data point is trained as a separate parameter.

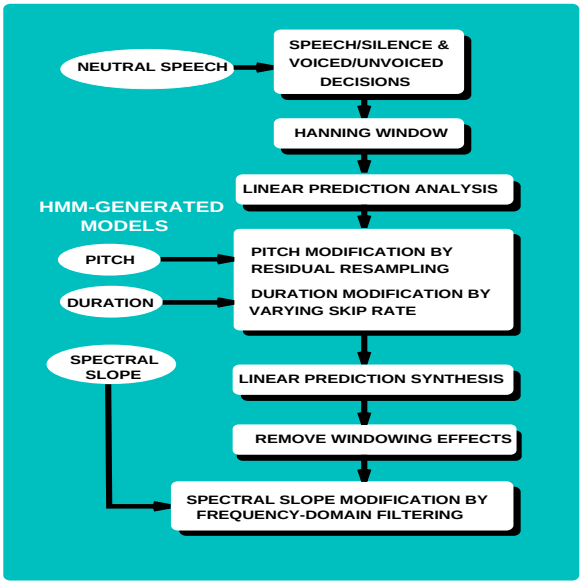

Figure 4: Speaking style modification using HMM-based models.

3.2 HMM-Generated Parameter Perturbation

The trained HMM models and the voiced duration transformation distributions described in the previous section are employed to generate duration scaling factors, pitch contours, and spectral slope mismatch perturbation sequences to be used for modifying neutral speech. In order to generate HMM-based pitch scaling vectors, two values should first be determined: the total number of observations to be produced by the HMM or the desired length of the pitch scaling profile, and the length of time spent in each state. The desired pitch profile length or voiced duration, is obtained by multiplying the duration of the neutral voiced speech by a scaling factor which accounts for the duration variation from neutral to stressed conditions. The duration scaling factor is randomly generated from the PMF of the voiced duration distribution. The time spent in a state is computed by sampling the state occupancy distribution associated with that state. Once the observations in a state are produced, they are ordered in either an ascending or descending order depending on the values generated in the previous state. The ordering is chosen so as to minimize the distance between the last data point in the previous state and the first data point in the current state. It is assumed here that the pitch profile is continuous. This assumption is valid since all words used for training consist of a single continuous voiced region with no unvoiced sections in between.

A spectral slope mismatch is generated by randomly selecting one sample from each of the 3 Gaussian distributions associated with the spectral slope HMM. These 3 points are then fitted to a 2^{nd} order polynomial which is the desired spectral slope mismatch.

The pitch scaling profiles and spectral mismatch perturbation response are then used to perturb neutral speech parameters as discussed in the following section.

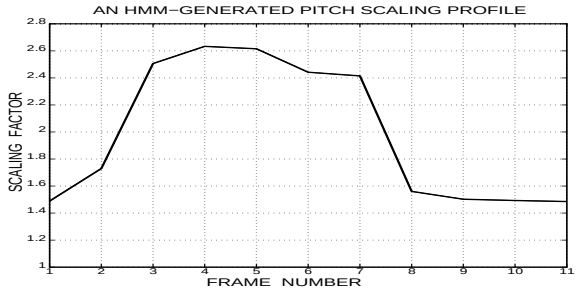

Figure 5: Pitch scaling profile generated by the 3-state HMM angry model.

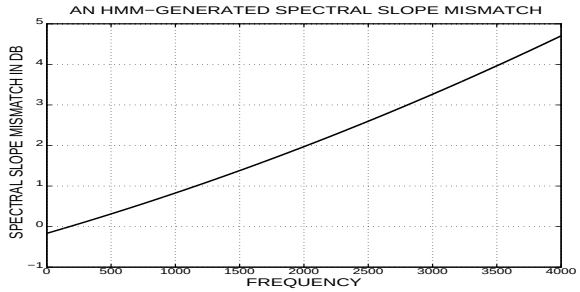

Figure 6: Spectral slope mismatch generated by the 1-state 3-parameter HMM for neutral-to-angry transformation.

4. SPEAKING STYLE MODIFICATION

Using the HMM-based stressed speech models from Sec. 3, a single overall algorithm is constructed which integrates pitch, duration, and spectral slope perturbation, as shown in Fig. 4, for generating stressed speech from neutral speech. Figures 5 and 6 show examples of pitch contour and spectral slope mismatch generated by the HMMs described in Sec. 3.2 and used for neutral-to-angry transformations as described below. Fig. 5 shows that the pitch of an input neutral word should be increased by a factor of 1.5 for the first and last frame, and by 2.6 for the 5^{th} frame. The spectral slope mismatch plot indicates that almost +5 dB of additional energy should be introduced into the neutral spectrum at 4 kHz (under angry conditions, high frequency content increases). The perturbation sequences for pitch and spectral slope agree with earlier statistical studies of speech under stress for the three styles under consideration [4].

Pitch and duration are then modified in the time domain within a linear prediction framework, while spectral slope modification is done in the frequency domain. Pitch is modified by re-sampling the linear prediction residual on a frame-by-frame basis according to the HMM-generated pitch profile. Duration modification is achieved in a pitch-synchronous manner by varying the rate at which the speech data is processed. Spectral tilt modification is achieved via frame-to-frame frequency filtering.

Next, the synthetic angry speech which resulted from neutral speech perturbation is presented to listeners in an effort to subjectively judge its stress content. This subjective listener evaluation was used previously for subjective assessment of speech under neutral and stressed conditions [1]. In this test, listeners heard a series of word pairs which consisted of either (a) an original neutral word and its neutral-to-angry modified word, or (b) an original angry word and the neutral-to-angry modified word. The word pairs were presented in a random order. Then, they were asked to pick 1 of 4 choices: (1) first word is more angry, (2) second word is more angry, (3) both words are equally angry, or (4) neither word is angry. The listener results are as follows. The neutral-to-angry modified speech was judged to sound more angry than the original neutral 97 out of the 112 open test occurances.

CONCLUSIONS

In this study, we have presented a novel approach for modeling the variation of speech parameters under stressed speech conditions based on Hidden Markov Models. Traditionally, HMMs have been used for speech recognition, but here they were employed for modeling and parameter generation in order to transform neutral speech into one of 3 stressed speaking conditions: angry, Lombard effect, and loud. The HMM model approach demonstrates that this approach is capable of capturing the wide variations due to stress in pitch, spectral slope, and duration. The generated HMM models could also be incorporated into stressed speech recognition systems to improve overall performance by generating stressed tokens for training as was suggested in [5].

References

[1] S. E. Bou-Ghazale, J.H.L. Hansen. A source generator based modeling framework for synthesis of speech under stress. In *Proc. IEEE ICASSP*, pp. 664–667, 1995.

[2] J. D. Ferguson. Variable duration models for speech. In *Proc. of the Symposium on the Applications of Hidden Markov to Text and Speech*, pp. 143–179, 1980.

[3] T. Fukada, Y. Komori, T. Aso, and Y. Ohora. A study on pitch pattern generation using HMM-based statistical information. In *ICSLP-94*, pp. 723–726, 1994.

[4] 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.

[5] J. H. L. Hansen, S. E. Bou-Ghazale. Robust Speech Recognition Training via Duration and Spectral-Based Stress Token Generation. *IEEE Trans. on Speech and Audio Proc.*, 3:415–421, September 1995.

[6] F. Itakura, S. Saito. Analysis synthesis telephony based upon the maximum likelyhood method. In Y. Kohasi, editor, *Reports of 6th Int. Cong. Acoust.* Tokyo, 1968. C-5-5,C17-20.

[7] 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, 1987.

[8] A. Ljolje, F. Fallside. Synthesis of natural sounding pitch contours in isolated utterances using hidden markov models. *IEEE Trans. on Acoustics, Speech, Signal Proc.*, pp. 1074–1080, Oct. 1986.