ANALYSIS OF MRATE, SHIMMER, JITTER, AND F_0 CONTOUR FEATURES ACROSS STRESS AND SPEAKING STYLE IN THE SUSAS DATABASE

Raymond E. Slyh[1] *W. Todd Nelson*[1] *Eric G. Hansen*[2]

[1] Air Force Research Laboratory, Human Effectiveness Directorate, Wright-Patterson AFB, Ohio
[2] Veridian, Dayton, Ohio

ABSTRACT

This paper highlights the results of an investigation of several features across the style classes of the "simulated" portion of the SUSAS database. The features considered here include a recently-introduced measure of speaking rate called mrate, measures of shimmer, measures of jitter, and features derived from fundamental frequency (F_0) contours. The F_0 contour features are the means of F_0 and ΔF_0 over the first, middle, and last thirds of the ordered set of voiced frames for each word. Mrate exhibits differences between the Fast, Neutral, and Slow styles and between the Loud, Neutral, and Soft styles. Shimmer and jitter exhibit differences that are similar to those of mrate; however, the shimmer and jitter differences are less consistent than the mrate differences across the speakers in the database. Several F_0 contour features exhibit differences between the Angry, Loud, Lombard, and Question styles and most of the other styles.

1. INTRODUCTION

The analysis and classification of speech from various speaking styles, stresses, and emotions is a problem that has generated considerable research in recent years [1–3]. One of the stressed speech databases often used in this research is the Speech Under Simulated and Actual Stress (SUSAS) database, particularly the "simulated" portion [1, 2]. Several acoustic features have been investigated to determine how they vary across the styles of the SUSAS database [1, 3]. This paper highlights the results of an investigation of additional acoustic features across the styles of the "simulated" portion of the database.

The "simulated" portion of the SUSAS database consists of utterances from nine male speakers in each of eleven stress/speaking-style classes. For the remainder of this paper, we simply refer to the stress and speaking-style classes as styles. The eleven styles are: Angry, Clear, Cond50, Cond70, Fast, Lombard, Loud, Neutral, Question, Slow, and Soft. The Cond50 and Cond70 styles consist of utterances from subjects engaged in tracking tasks; the Cond50 style is the result of a medium workload condition, while the Cond70 style is the result of a high workload condition. The Lombard style consists of utterances from subjects listening to pink noise presented binaurally through headphones at a level of 85 dBA. Each speaker uttered the same 70 isolated words for each style; the 70 words consisted of two repetitions of a 35-word list.

The features considered here include a recently-introduced measure of speaking rate called mrate [4]; two measures of shimmer, Shim and ShdB; two measures of jitter, Jita and Jitt; and features derived from fundamental frequency (F_0) contours. Mrate combines multiple measures of speaking rate to yield a single measure of speaking rate in syllables per second. Shim measures the period-to-period (PTP) variability of the peak-to-peak amplitude and is expressed as a percentage:

$$\text{Shim} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i},$$

where the A_i are the extracted peak-to-peak amplitude data, and N is the number of extracted pitch periods. ShdB measures the PTP variability of the peak-to-peak amplitude in decibels:

$$\text{ShdB} = \frac{1}{N-1}\sum_{i=1}^{N-1}|20\log(A_{i+1}/A_i)|.$$

Jita measures the PTP variability of the pitch period, T_0, in μsec:

$$\text{Jita} = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|T_{0_i} - T_{0_{i+1}}\right|,$$

where the T_{0_i} are the extracted pitch period lengths. Jitt measures the relative PTP variability of T_0 and is expressed as a percentage:

$$\text{Jitt} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|T_{0_i} - T_{0_{i+1}}\right|}{\frac{1}{N}\sum_{i=1}^{N} T_{0_i}}.$$

The F_0 contour features are the means of F_0 and ΔF_0 over the first, middle, and last thirds of the ordered set of voiced frames for each word. In the remainder of the paper, $\overline{F_0(i)}$ and $\overline{\Delta F_0(i)}$ denote the means of F_0 and ΔF_0, respectively, over the i th third of the ordered set of voiced frames for a word.

For each speaker and feature pair, we used a Multivariate Analysis of Variance (MANOVA) [5] to determine if any statistically significant differences (SSDs) existed between the feature means for the various styles for that speaker. If the MANOVA indicated that an SSD existed, then we conducted pairwise comparisons to determine which style pairs differed. We considered a feature as exhibiting consistent differences across speakers for two styles if 75% or more (*i.e.*, seven to nine) of the speakers exhibited SSDs for that style pair. Finally, we conducted style classification experiments using some of the features exhibiting consistent differences between various styles.

The organization of the paper is as follows. The next section outlines the methods used to extract the various features. Section 3

highlights the statistical analysis of the features. Section 4 highlights the results of a style classification experiment using some of the features exhibiting consistent differences across speakers. Finally, Section 5 presents a discussion and the conclusions.

2. FEATURE EXTRACTION

We extracted the features considered here using three main procedures. First, we extracted the mrate feature using the software[1] of Morgan and Fosler-Lussier. Second, we extracted the shimmer and jitter features using the Multidimensional Voice Program (MDVP) from the Kay Elemetrics Corporation [6]. Finally, we extracted the F_0 contour features using the output of the *get_f0* program available in the Entropic Signal Processing System (ESPS) from the Entropic Research Laboratory. The *get_f0* program is an implementation of the pitch-tracking algorithm described in [7].

To extract the jitter and shimmer features, we used the "running speech mode" of the MDVP. This mode requires the speech signal to have a sampling frequency of at least 25 kHz, so we upsampled the SUSAS utterances from their 8 kHz sampling rate to 25 kHz. The MDVP would not process some of the upsampled SUSAS utterances with large amplitudes, so we reduced the amplitude of all of the upsampled signals by one half. For several of the SUSAS utterances, the MDVP did not find enough pitch periods to analyze the speech. The styles and speaker pairs for which the MDVP did not successfully analyze five or more utterances are: (1) the Soft style for speakers B1, B2, G2, and G3; (2) the Fast style for speakers B1, G2, G3, N1, and N2; (3) the Angry style for speaker N1; and (4) the Question style for speaker G2. The missing feature values for these words impacted our choice of analysis methods (see Section 3).

We calculated the F_0 contour features in the following manner. For each word, we used the *get_f0* program to determine F_0 values and voicing information for speech frames spaced 10 msec apart. We computed both ΔF_0 and $\Delta^2 F_0$ values as follows: $\Delta F_{0_i} = F_{0_i} - F_{0_{i-1}}$ and $\Delta^2 F_{0_i} = \Delta F_{0_i} - \Delta F_{0_{i-1}}$, where F_{0_i} is the F_0 value for the i th frame. The *get_f0* program sets F_0 to zero for unvoiced frames, so the ΔF_0 value for the first frame in each voiced segment was invalid. Likewise, the $\Delta^2 F_0$ values for the first two frames in a voiced segment were invalid. We extracted the F_0, ΔF_0, and $\Delta^2 F_0$ values from the voiced frames of a word with the exception of the first two frames from each voiced segment (eliminating a small amount of valid F_0 and ΔF_0 data). In performing this extraction of valid voiced frames, we kept the relative time-ordering of the extracted frames intact. We partitioned the ordered set of extracted voiced frames for a word into thirds. Groups N_1, N_2, and N_3 consisted of the frames from the first, middle, and last thirds of the ordered set of extracted frames, respectively. We calculated $\overline{F_0(i)}$ as the mean of the F_0 values over group N_i, $\overline{\Delta F_0(i)}$ as the mean of the ΔF_0 values over group N_i, and $\overline{\Delta^2 F_0(i)}$ as the mean of the $\Delta^2 F_0$ values over group N_i. The $\overline{\Delta^2 F_0(i)}$ features are not considered here as they exhibited less consistent differences across speaker than did the other F_0 contour features.

3. FEATURE ANALYSIS

For the shimmer and jitter features, only three words (*degree*, *freeze*, and *oh*) were successfully analyzed across all of the speakers, styles, and repetitions (see Section 2). To conduct an analysis of variance (ANOVA) across these three factors, we would have had to estimate the missing data for several words or ignore the data for those words that had incomplete data across the three factors. Instead, we chose to analyze the data for each speaker separately, thereby taking into account considerably more of the available data. For each speaker/feature pair, we conducted a two-way (11 styles $\times$ 2 repetitions) dependent-samples (*i.e.*, within-subjects) MANOVA followed by paired comparisons. Here, words played the role that subjects normally play in a two-way within-subjects design. We chose to use a MANOVA rather than an ANOVA on the recommendation of [5]. This recommendation centers around the fact that the dependent-samples ANOVA is sensitive to violations of its sphericity assumption on the data, while the dependent-samples MANOVA has no sphericity requirement on the data [5].

We tested the main effects and interactions using a significance level of 0.05. For the features considered here, every speaker/feature pair exhibited a statistically significant style effect at the $p < 0.05$ level. Some of the speaker/feature pairs exhibited statistically significant repetition effects at the $p < 0.05$ level. The repetition effects for the F_0 contour features were sometimes quite large (see Section 3.4). All of the observed statistically significant style $\times$ repetition effects occurred for F_0 contour features.

For the pairwise style comparisons, we used dependent-samples t-tests with the Bonferroni procedure to control the familywise error rate [5]. The Bonferroni procedure keeps the familywise error rate below a level of α by using a significance level of $\alpha_P = \alpha/C$ for each pairwise comparison, where C is the number of pairwise comparisons to be made. For the 11 SUSAS styles, there are $C = 55$ possible pairwise comparisons. Thus, to keep the familywise error rate below $\alpha = 0.05$, we used $\alpha_P = 0.05/55$ as the significance level for each pairwise comparison. In the following subsections, we denote the mean of the feature values for a particular style by the feature name with the style as a subscript.

3.1. Mrate

Table 1 shows the style pairs with SSDs in mrate at the $p < \alpha_P$ level for seven or more of the nine speakers. For a given style pair in Table 1, the "Speaker Exceptions" column lists the speakers (if any) that did not exhibit an SSD for the pair. Some of the consistent SSDs are: (1) mrate$_{Neutral}$ is greater than mrate$_{Slow}$ by 0.39–1.31 syllables/sec for all of the speakers, (2) mrate$_{Fast}$ is greater than mrate$_{Neutral}$ by 0.31–1.14 syllables/sec for all of the speakers except N2, (3) mrate$_{Neutral}$ is greater than mrate$_{Loud}$ by 0.36–0.71 syllables/sec for all of the speakers except B2 and N3, and (4) mrate$_{Soft}$ is greater than mrate$_{Loud}$ by 0.57–1.08 syllables/sec for all of the speakers except N3. Thus, the mrate values increase as one progresses from the Slow style through the Neutral style to the Fast style. The mrate values generally decrease as one progresses from the Soft or Neutral styles to the Loud style.

The speakers exhibit standard deviations in the mrate values for the styles ranging from 0.32 to 0.94 syllables/sec, which are on the order of the differences in the means between the styles. As an example, speaker B1 has standard deviations of 0.40, 0.66, and 0.73 syllables/sec for the Slow, Neutral, and Fast styles, respectively. The means are mrate$_{Slow}$ = 2.46, mrate$_{Neutral}$ = 3.24, and mrate$_{Fast}$ = 3.55. Thus, the mean plus the standard deviation for the Slow style (2.86) is larger than the mean minus the standard deviation for the Fast style (2.82). These results indicate a substantial overlap between the three styles in terms of mrate for speaker B1.

[1] http://www.icsi.berkeley.edu/ftp/global/pub/speech/morgan/

Table 1: Style Pairs with SSDs in Mrate for 7–9 Speakers

Style Pair	Speaker Exceptions	Style Pair	Speaker Exceptions
Angry, Fast	N1	Fast, Slow	
Angry, Soft	G2 N1	Lomb., Neut.	G3
Clear, Fast		Lomb., Quest.	G3 N3
Cond50, Fast	B2 N2	Lomb., Soft	G3
Cond50, Lomb.	G3	Loud, Neut.	B2 N3
Cond50, Slow	N2	Loud, Quest.	N3
Cond70, Fast	B2 N2	Loud, Slow	G2 N2
Cond70, Lomb.	G2	Loud, Soft	N3
Cond70, Slow	G2	Neut., Slow	
Fast, Lomb.		Quest., Slow	
Fast, Loud		Slow, Soft	
Fast, Neut.	N2		

Table 2: Style Pairs with SSDs in Shim or ShdB for 7–9 Speakers

Style Pair	Shim Speaker Exceptions	ShdB Speaker Exceptions
Angry, Fast	B1 N1	B1 N1
Clear, Fast	B3 N1	B3 N1
Cond50, Fast	B2 G1	—
Cond50, Lombard	B1 B3	B1 B3
Cond70, Loud	G2	G2 N3
Cond70, Slow	G2 N2	G2 N2
Fast, Lombard		B1
Fast, Loud		
Fast, Slow		
Lombard, Soft		B1
Loud, Neutral		B1 N3
Loud, Soft		N3
Neutral, Slow	—	G2 N2
Question, Slow	B1	B1 N3
Slow, Soft		

3.2. Shim and ShdB

Table 2 shows the style pairs with SSDs in Shim or ShdB at the $p < \alpha_P$ level for seven or more of the nine speakers. A dash in a feature's exception column for a style pair indicates that the feature did not exhibit SSDs for that style pair for seven or more speakers. For Shim, all of the speakers exhibit the following SSDs: (1) Shim_{Fast} is greater than Shim_{Slow} by 3.60–10.33%, (2) $\text{Shim}_{Neutral}$ is greater than Shim_{Loud} by 1.59–5.98%, and (3) Shim_{Soft} is greater than Shim_{Loud} by 1.95–11.01%. For ShdB, some of the consistent SSDs are: (1) ShdB_{Fast} is greater than ShdB_{Slow} by 0.40–1.23 dB for all of the speakers, (2) $\text{ShdB}_{Neutral}$ is greater than ShdB_{Slow} by 0.28–0.65 dB for all of the speakers except G2 and N2, (3) $\text{ShdB}_{Neutral}$ is greater than ShdB_{Loud} by 0.22–0.52 dB for all of the speakers except B1 and N3, and (4) ShdB_{Soft} is greater than ShdB_{Loud} by 0.22–0.99 dB for all of the speakers except N3. Thus, both Shim and ShdB increase as one progresses from the Slow style to the Fast style. Additionally, ShdB generally increases as one progresses from the Slow style to the Neutral style. Shim and ShdB generally decrease as one progresses from the Soft or Neutral styles to the Loud style. The standard deviations for Shim range from 1.11 to 6.53% across style and speaker. The standard deviations for ShdB range from 0.14 to 0.66 dB across style and speaker.

3.3. Jita and Jitt

Jita and Jitt exhibit fewer consistent SSDs between style pairs than do mrate, Shim, or ShdB. For all of the speakers except B1 and N2, (1) Jita_{Fast} is greater than Jita_{Slow} by 75–210 μsec and (2) Jitt_{Fast} is greater than Jitt_{Slow} by 1.19–2.84%. The standard deviations for Jita range from 19.83 to 418.49 μsec across style and speaker, while the standard deviations for Jitt range from 0.40 to 4.31%.

3.4. F_0 Contour Features

Some of the consistent style differences for the F_0 contour features are:

- $\overline{F_0(1)}$: between the Angry and Loud styles and the other styles
- $\overline{\Delta F_0(1)}$: between the Loud style and the other styles except for the Angry style
- $\overline{F_0(2)}$: between the Loud style and the other styles, between the Angry style and the other styles except for the Question style, and between the Lombard style and the other styles except for the Clear and Question styles
- $\overline{\Delta F_0(2)}$: between the Question style and the other styles
- $\overline{F_0(3)}$: between the Question style and the other styles
- $\overline{\Delta F_0(3)}$: between the Angry and Loud styles and the other styles

As previously indicated, some of the F_0 contour features exhibit large repetition effects or style $\times$ repetition interactions for some of the speakers. We consider two of these cases here—namely, repetition effects in $\overline{F_0(3)}$ for speakers B1 and N2. In the next section, we show how these two repetition effects impact style classification results.

The repetition effect in $\overline{F_0(3)}$ for speaker B1 is mostly between the repetitions of the Loud style. For the first word list of the Loud style, the mean and standard deviation of $\overline{F_0(3)}$ are 119 Hz and 22 Hz, respectively. For the second word list of the Loud style, the mean and standard deviation are 214 Hz and 16 Hz, respectively. The means of the two word lists for the Question style are 217 Hz and 211 Hz, while the means for the styles other than the Loud and Question styles are all below 131 Hz. Thus, the $\overline{F_0(3)}$ values for the second word list of the Loud style are in the range of those for the Question style, while the $\overline{F_0(3)}$ values for the first word list of the Loud style are in the range of the styles other than the Loud and Question styles.

The repetition effect in $\overline{F_0(3)}$ for speaker N2 consists of repetition effects for the Neutral, Question, Slow, and Soft styles. The means of the first word lists for these four styles are: (1) Neutral: 127 Hz, (2) Question: 208 Hz, (3) Slow: 107 Hz, and (4) Soft: 194 Hz. The means of the second word lists for these four styles are: (1) Neutral: 86 Hz, (2) Question: 125 Hz, (3) Slow: 81 Hz, and (4) Soft: 92 Hz. Thus, the $\overline{F_0(3)}$ values for the second list of the Question style are in the range of those for the first list of the Neutral style. The $\overline{F_0(3)}$ values for the second lists of the Neutral, Slow, and Soft styles are in the range of those for the Clear style.

4. STYLE CLASSIFICATION

This section highlights the results of a style classification experiment using the mrate, $\overline{F_0(1)}$, $\overline{F_0(2)}$, $\overline{F_0(3)}$, $\overline{\Delta F_0(1)}$, $\overline{\Delta F_0(2)}$, and $\overline{\Delta F_0(3)}$ features. For this experiment, we grouped the styles as in [3]. The groups are: (S1) the Angry and Loud styles; (S2) the Cond50, Cond70, Neutral, and Soft styles; (S3) the Fast style; (S4)

Table 3: Correct Classification Rates for the Style Groups: (S1) Angry and Loud; (S2) Cond50, Cond70, Neutral, and Soft; (S3) Fast; (S4) Question; (S5) Slow; (S6) Clear; and (S7) Lombard

Spkr	S1	S2	S3	S4	S5	S6	S7
B1	8.6	40.7	37.1	97.1	48.6	54.3	74.3
B2	92.9	30.0	51.4	85.7	60.0	37.1	80.0
B3	94.3	18.6	62.9	91.4	74.3	40.0	82.9
G1	98.6	80.7	57.1	97.1	77.1	34.3	71.4
G2	84.3	33.6	65.7	68.6	40.0	51.4	65.7
G3	100.0	37.1	14.3	88.6	82.9	42.9	88.6
N1	77.1	73.4	8.6	71.4	57.1	48.6	91.4
N2	41.4	28.6	14.3	42.9	20.0	51.4	74.3
N3	95.7	59.3	60.0	100.0	62.9	40.0	100.0

the Question style; (S5) the Slow style; (S6) the Clear style; and (S7) the Lombard style. Using the first word list of each style group, we trained speaker-dependent Maximum Likelihood classifiers (assuming multivariate Gaussian densities for the features). We tested the classifiers using the second word list for each style group. Table 3 shows the classification rates for the seven groups for each speaker.

Group S1 generally shows good results except for speakers B1 and N2. For speaker B1, the repetition effect in $\overline{F_0(3)}$ for the Loud style (see Section 3.4) causes almost all of the words from the second word list for the Loud style to be classified as belonging to the Question style. Thus, group S1 is confused with group S4 47.1% of the time for speaker B1. Speaker N2 exhibits repetition effects for all of the F_0 contour features that we consider here. In almost every case, the repetition effects result in the feature means of the second word lists for the Angry and Loud styles moving closer to those of the Lombard style. These effects cause group S1 to be confused with group S7 38.6% of the time for speaker N2.

Group S4 shows good results except for speaker N2. For N2, the repetition effect in $\overline{F_0(3)}$ for the Question style (see Section 3.4) leads to group S4 being confused with group S2 37.1% of the time.

Group S7 shows good results except for speaker G2. For speaker G2, group S7 is confused with group S1 28.6% of the time.

The S2, S3, S5, and S6 groups do not show consistent results across speakers. For speaker N2, the repetition effects in $\overline{F_0(3)}$ for the Neutral, Slow, and Soft styles (see Section 3.4) lead to groups S2 and S5 being confused with group S6. Group S2 is confused with group S6 47.9% of the time, while group S5 is confused with group S6 65.7% of the time. The large degree of overlap between various styles in terms of mrate also leads to confusions among these groups.

5. DISCUSSION AND CONCLUSIONS

A number of features were investigated and found to exhibit consistent differences across speakers for various style pairs. Despite the consistent differences between styles, many of these features also exhibit considerable variability within styles leading to large degrees of overlap between the feature distributions for various styles. Mean mrate values increase as one progresses from the Slow style through the Neutral style to the Fast style and decrease as one progresses from the Soft or Neutral styles to the Loud style. At the same time, the individual mrate values generally exhibit large degrees of overlap between the Slow, Neutral, and Fast styles and between the Soft, Neutral, and Loud styles. The shimmer and jitter features exhibit differences that are similar to those of mrate; however, the shimmer and jitter differences are less consistent than the mrate differences across the speakers in the database. The F_0 contour features considered here exhibit differences between the Angry, Loud, Lombard, and Question styles and most of the other styles. For most speakers, the mrate and F_0 contour features combine to give good classification rates for the Question style, the Lombard style, and the Angry and Loud style group. However, some of the speakers (particularly B1 and N2) exhibit large repetition effects in the F_0 contour features that adversely affect the classification rates.

There are two additional points to consider from this work. First, the MANOVA framework was useful in identifying repetition effects and style $\times$ repetition interactions in the F_0 contour features that adversely affected style classification results. We are continuing to use the MANOVA framework to investigate additional features across the styles and speakers of the database. Second, it is unclear whether the repetition effects for the F_0 contour features affect listeners' perception of the speaking styles, and it is unclear whether the styles that have large degrees of overlap in terms of mrate, shimmer, and jitter are perceived by listeners as having large degrees of overlap. To address these perceptual aspects, we are conducting listening tests to determine how well humans classify the styles in the database.

6. ACKNOWLEDGEMENTS

The authors wish to thank Dr. Timothy Anderson of the Air Force Research Laboratory (AFRL) for valuable discussions on stressed speech analysis and classification. Special thanks go to Jonathan Pyles and David Russo, both summer employees of AFRL, for extracting the MDVP features from the SUSAS database. Finally, thanks go to Prof. Nelson Morgan and Eric Fosler-Lussier, both of the International Computer Science Institute at the University of California at Berkeley, for making the mrate code available.

7. REFERENCES

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

[2] J. H. L. Hansen and S. E. Bou-Ghazale, "Getting started with SUSAS: A speech under simulated and actual stress database," in *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, (Rhodes, Greece), pp. 1743–1746, September 1997.

[3] J. H. L. Hansen and B. D. Womack, "Feature analysis and neural network-based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 307–313, July 1996.

[4] N. Morgan and E. Fosler-Lussier, "Combining multiple estimators of speaking rate," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Seattle, WA), pp. II–729–732, May 1998.

[5] S. E. Maxwell and H. D. Delaney, *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. Pacific Grove, CA: Brooks/Cole, 1990.

[6] D. D. Deliyski, "Acoustic model and evaluation of pathological voice production," in *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, (Berlin, Germany), pp. 1969–1972, September 1993.

[7] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), New York: Elsevier, 1995.