On the robustness of overall F0-only modifications to the perception of emotions in speech

Murtaza Bulut[a] and Shrikanth Narayanan
Signal Analysis and Interpretation Laboratory, http://sail.usc.edu, Electrical Engineering Department, University of Southern California, Los Angeles, California 90089

(Received 26 October 2006; revised 24 March 2008; accepted 24 March 2008)

Emotional information in speech is commonly described in terms of prosody features such as F0, duration, and energy. In this paper, the focus is on how F0 characteristics can be used to effectively parametrize emotional quality in speech signals. Using an analysis-by-synthesis approach, F0 mean, range, and shape properties of emotional utterances are systematically modified. The results show the aspects of the F0 parameter that can be modified without causing any significant changes in the perception of emotions. To model this behavior the concept of emotional regions is introduced. Emotional regions represent the variability present in the emotional speech and provide a new procedure for studying speech cues for judgments of emotion. The method is applied to F0 but can be also used on other aspects of prosody such as duration or loudness. Statistical analysis of the factors affecting the emotional regions, and discussion of the effects of F0 modifications on the emotion and speech quality perception are also presented. The results show that F0 range is more important than F0 mean for emotion expression.
 [DOI: 10.1121/1.2909562]

PACS number(s): 43.72.Ar [DOS] Pages: 4547–4558

I. INTRODUCTION

Studies of emotional speech have shown that emotion change can be associated with changes in the prosodic and spectral characteristics of speech signals (Bulut *et al.*, 2005, 2002; Burkhardt and Sendlmeier, 2000; Cahn, 1990; Montero *et al.*, 1999). The focus has been mainly on prosody parameters, such as F0, duration, and energy. Among these, significant attention has been paid to F0 contour modulations occurring as a result of emotion change (Cowie *et al.*, 2001; Murray and Arnott, 1993; Scherer, 2003).

Acoustic analyses of angry or happy speech show that, in general, their F0 mean, median, range, and variance values are larger than their neutral speech counterparts, which are larger than sad emotion F0 values (Davitz, 1964; Iida *et al.*, 2003; Murray and Arnott, 1993). The F0 contours of happy and angry speech, in most cases, are more variable than neutral speech, showing fast and irregular up and down movements, while the sad speech F0 contours show smaller variation and downward inflections (Davitz, 1964; Murray and Arnott, 1993). Although these findings are fairly consistent across different studies, differences are not uncommon. For instance, in Yildirim *et al.* (2004) sad speech had a higher F0 mean than neutral speech.

Despite having a powerful descriptive value, the aforementioned technique for studying emotions has several limitations. For example, its implementation in emotional speech synthesis is limited (Cowie *et al.*, 2001) because it does not specifically account for the variability present in the natural speech (Braun *et al.*, 2006; Chu *et al.*, 2006; Pell, 2001). In the traditional analysis, an emotional utterance is represented as a point in the parameter space. We suggest a new model where each utterance is represented by an "*emotional region*" in the parameter space. The proposed model is a new procedure for studying speech cues for judgments of emotion. In this paper, the method was applied to F0 but it can also be used on other aspects of prosody such as duration or loudness.

In this paper, we show how F0 mean, range, and shape characteristics can vary in emotional utterances, and how these variations can be modeled using the emotional regions. Statistical analyses of the factors that cause the variability are presented. In addition, the effects of different F0 modifications on emotional content perception are also investigated. These results show that F0 range is more important than F0 mean for emotional expression.

The concept of the variability of prosodic patterns was studied in a database composed of two repetitions of 1000 sentences recorded with six months separation by Chu *et al.* (2006). The results showed wide variations in F0 values, sometimes corresponding to 50% of the dynamical range of the speaker. In an another study (Braun *et al.*, 2006) iterative mimicry was employed to observe whether F0 contours converge to specific English intonation patterns, referred to as "attractors." It was only after several iterations that F0 branching (i.e., clustering) patterns were seen. However, even then the variability of F0 contours was noticeable. This was due to the fact that "human variability places a lower limit on the width of the branches" (Braun *et al.*, 2006).

In this paper the concept of emotional regions is introduced to model the variability in the F0 characteristics of emotional utterances. A model based on F0 mean, range, and standard deviation statistics is proposed. Following an analysis-by-synthesis approach it is shown that the proposed

[a]Author to whom correspondence should be addressed. Electronic mail: murtaza.bulut@philips.com

model gives reliable estimation of how F0 contours of emotional utterances can be modified without significantly affecting the perceived emotional content and speech quality. This representation is helpful to better assess the role of F0 contours in emotion perception. Also, when applied together with the F0 generation models such as ToBI (Silverman *et al.*, 1992) or Tilt (Taylor, 2000), it can be used to better predict the intonation events in emotional speech synthesis.

Emotion perception is a result of the interplay between acoustical, lexical, and environmental factors (Traunmuller, 2005). These factors can be expected to have an effect on the emotional regions. We show how the speaker and utterance characteristics affect the emotional regions, and analyze the interaction between different factors using statistical methods.

The effects of modifying F0 contour shape, F0 range, and voice quality characteristics of emotional utterances (in German) were statistically analyzed by Ladd *et al.* (1985) for arousal related (relaxed/aroused, open/deceitful, annoyed/content, insecure/arrogant, and indifferent/involved) and cognition related (emphasis, cooperativeness, contradiction, surprise, and reproach) emotions. The results showed that *text* (i.e., sentence content) had a significant effect on listener judgment. Similarly, the *speaker* factor (i.e., who uttered the utterance) also had significant effect for all emotion categories, except *arrogant*. The results also showed that modifying F0 range had a significant (and *continuous*) effect on emotion perception, especially on speaker arousal. The effects of contour changes were less prominent than range modifications. Similarly, in this paper, we also analyze the effects of F0 range and contour modifications, and also consider the *sentence* and *speaker* as independent factors. However, our focus, in contrast, is on how F0 acoustic features interact with *sentence*, *speaker*, and *emotion* factors in influencing the perception of emotion and speech quality. We use *angry*, *happy*, *sad*, and *neutral* labels, which are a subset of the emotional labels suggested by Ekman and Friesen (1977), to describe emotions.

In this paper, the results were also analyzed from the emotional speech synthesis perspective (Burkhardt and Sendlmeier (2000); Cahn (1990); Raux and Black (2003); Schroder (2001)). It was observed that F0 modification caused the perception of sad and neutral emotions to increase, and angry and happy emotions to decrease. The effects of F0 range modifications were *continuous* (Ladd *et al.* (1985)) and more significant than F0 mean modifications. F0 contour shape modifications were also effective but only when performed in large semitone scales. It was also observed that the listeners were still able to perceive the emotions in a manner similar to that of the unmodified natural utterances even when the speech quality was distorted.

In the next sections we first describe the performed F0 mean, range, and shape modifications (Sec. II) and how they were evaluated (Sec. III). The concept of emotional regions is introduced in Sec. IV and the statistical analyses results are presented in Sec. V. The effects of F0 modifications on emotional content are presented in Sec. VI. The discussion and conclusion follow in Sec. VII and Sec. VIII, respectively.

II. DATA PREPARATION

In this section the emotional data collection and the F0 modifications are explained.

A. Data collection

Two sentences, "*She told me what you did*." (sentence 1) and "*This hat makes me look like an aardvark*." (sentence 2) were recorded by a female speaker (speaker 1) and a male speaker (speaker 2). Both speakers were in their late 20s. Speaker 1 had some professional acting experience, while speaker 2 did not.

The speakers were instructed to utter the two sentences in *angry*, *happy*, *sad*, and *neutral* (i.e., no particular emotion) emotion styles, resulting in a total of 16 utterances (see Fig. 1). However, no specific instructions were given on how the emotions should be expressed. In other words, the interpretation and expression of emotions was left to the speakers themselves. The speech was recorded in a quiet room at 48 kHz sampling rate using unidirectional head-worn dynamic Shure brand (model SM10) microphones. Later the speech was down sampled to 16 kHz. Listening tests were conducted, afterwards, to evaluate the success of emotion production. The results showed that human listeners were able to correctly identify (with approximately 80% success on average) the emotions expressed by the speakers.

B. F0 modifications

Several modifications manipulating the mean, range, and shape of the natural F0 contours were applied to all recorded emotional utterances (which will be also referred to as *original* utterances). The F0 mean, range, and shape modifications were performed using the Time Domain Pitch Synchronous Overlap and Addition (TD-PSOLA) algorithm (Moulines and Charpentier, 1990) as implemented in the Praat software (Boersma and Weenink, 2007).

The applied modifications can be categorized into three groups: Mean, range, and stylization modifications (summarized in Table I).

Modifications in F0 mean: The mean was modified by shifting the F0 contour up or down. The following modifications were applied: (1) Increasing/decreasing the original F0 mean by 10%, 15%, 25%, and 50%, (2) Making the F0 mean equal to 50, 100, 150, 200, 250, and 300 (Hz).

Modifications in F0 range: The range was modified by multiplying the F0 contour with a constant and then shifting the contour up or down so that the mean will be the same as the original mean value. The following modifications were applied: (1) Scaling the range by 0.5, 0.75, 1.5, and 2, (2) Making the F0 range equal to 10, 30, 50, 80, 110, and 150 (Hz).

Stylization modifications: The shape of the F0 contour of the utterances was altered by stylizing the F0 contour. The following modifications were applied: Stylizing the F0 contour by a 2, 5, 10, 15, and 40 semitone frequency resolution.

Stylization of the F0 contour was performed using the Praat software. The logic behind the stylization algorithm is to try to represent the F0 contour using linear segments. The length of the linear segments was determined by the fre-

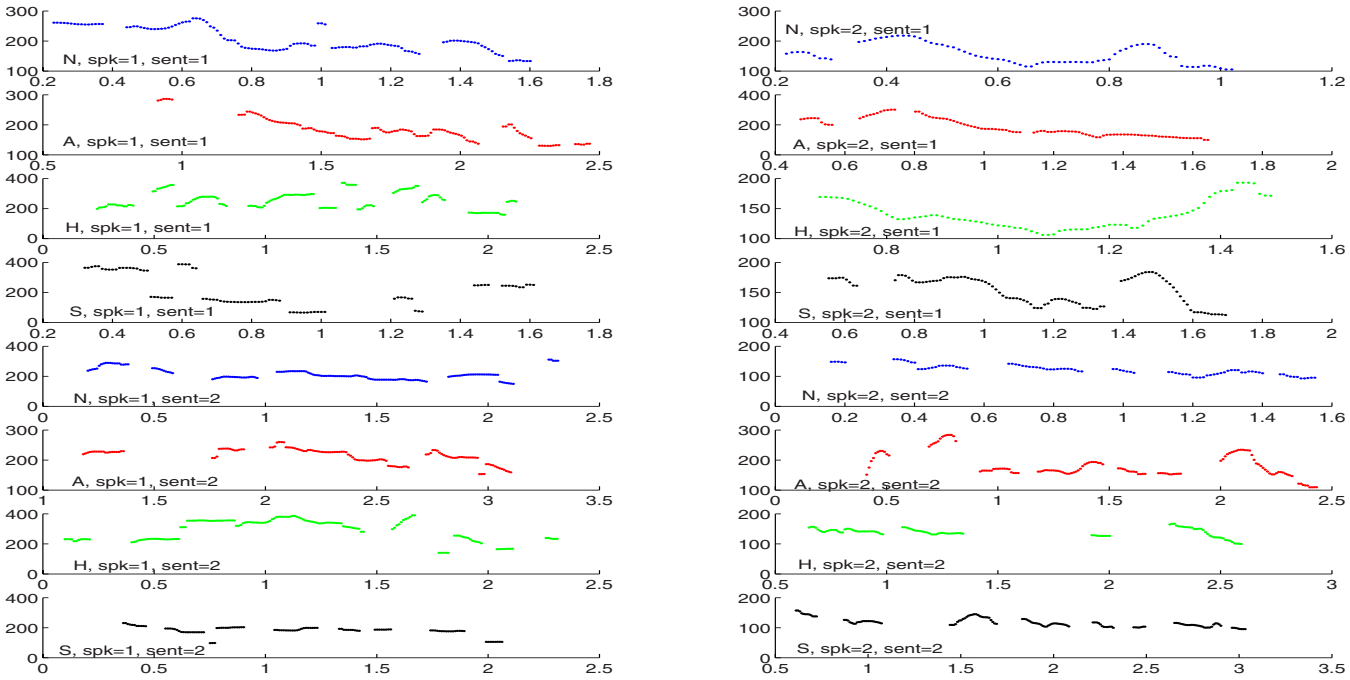

FIG. 1. (Color online) F0 contours of all 16 utterances that were recorded. *H*, *A*, *N*, *S*, *spk*, and *sent* denote happy, angry, neutral, sad, speaker, and sentence, respectively.

quency resolution component. For instance, while a 2 semitone resolution corresponds to fairly short linear segments, thus preserving the general contour shape, 40 semitone resolution may cause the whole utterance F0 contour to be a line (see Fig. 2 for an example).

As a result of applying the aforementioned modifications, 29 utterances, all having exactly the same duration as the original utterance but different F0 contours, were resynthesized for each of the original (i.e., recorded, natural) utterances. In total, including the original utterances, there were $(30\times 16=)$ 480 utterances.

III. LISTENING TESTS

All natural and resynthesized utterances were evaluated by listening experiments with naive listeners that included both native and non-native American English speakers. Before evaluation, all speech files were normalized so that the maximum digitized waveform amplitude was 1. In the listening tests—conducted in a quiet room, using headphones and with a single rater at a time—first the speech file was presented and then the raters were asked to choose among the following options: *Happy*, *angry*, *sad*, *neutral*, and *other*. The raters were particularly instructed to choose *other* if their choice of emotion was not listed or if they could not decide on the emotional content, or if the speech sounded to them as a mixture of several emotions. They were allowed to listen to each utterance as many times as they liked before making their decision. After the raters had chosen the emotion, they were asked to rate the naturalness (i.e., speech quality) of the utterance on a scale from 1 to 5, with 5 cor-

TABLE I. Summary of the performed F0 contour modifications. The values for mean and range are in Hz and the values for stylization are in semitone.

F0	Mean	Range	Stylization
Increase	m1: +10%	r3: +50%	
	m2: +15%	r4: +100%	
	m3: +25%		
	m4: +50%		
Decrease	m5: −10%	r1: −50%	
	m6: −15%	r2: −25%	
	m7: −25%		
	m8: −50%		
Set value	m9: =50	r5:= 10	s1:= 2
	m10:= 100	r6:= 30	s2:= 5
	m11:= 150	r7:= 50	s3:= 10
	m12:= 200	r8:= 80	s4:= 15
	m13:= 250	r9:= 110	s5:= 40
	m14:= 300	r10:= 150	

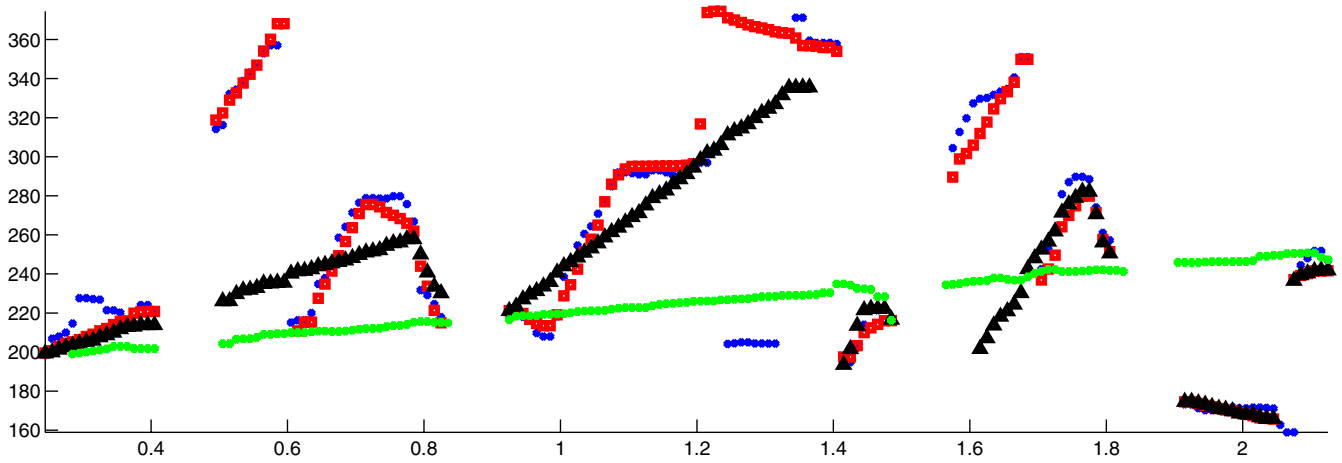

FIG. 2. (Color online) Stylization example for the happy utterance, speaker=1, sentence=1. Circles=original F0 contour, Squares=2 semitones stylization, Triangles=10 semitones stylization, Dots=40 semitones stylization.

responding to the most natural. They were specifically instructed to give low values if the speech was perceived to be different from natural human speech in terms of quality. Again, the raters were able to listen to the speech as many times as they liked. The files were presented in a different random order for each rater.

From the preliminary listening tests it was observed that long lasting tests were tiring for listeners (which may negatively affect their judgment abilities). Because of that, the average test time was limited to 20 min. per set. In order to limit the time of any single test to around 20 min., the test set was divided into ten groups of 48 utterances, each consisting or three variations of the 16 original utterances (which were chosen randomly). After the completion of a set, listeners were given the opportunity to rest (or to continue some other time), or to continue with a different test set.

The average number of raters per set was 9.2. In total, there were 14 different people that participated. Of these, seven people (three female, and four male listeners) evaluated all utterances. Most of the raters were graduate students in their mid to late 20s.

IV. EMOTIONAL REGIONS IN F0 MEAN-RANGE SPACE

One of the basic characteristics of natural speech is its variability (Braun *et al.*, 2006; Chu *et al.*, 2006). In order to generate models for speech production, synthesis, and perception this variability should be accounted for appropriately. In this section, we show examples of the variability present in emotional speech and propose a model to parametrize it.

For each of the resynthesized utterances, the F0 contour was calculated using the Praat software. After removing the outliers and smoothing using a median filter of length 3, F0 mean, F0 range [=(0.975 quantile)-(0.025 quantile)], and F0 standard deviation (std) statistics were calculated.

Based on the results of the listening tests, all resynthesized utterances were assigned an emotional label using majority voting. Then, each of these utterances was grouped together with its original version (i.e., the utterance from which it was resynthesized) only if its emotion was the same as the emotion of the original utterance. As a result, 16 (=2 speakers×2 sentences×4 emotions) groups (one for each original utterance) were generated. The utterances in these groups were used to construct the emotional regions.

We introduce the idea of emotional regions to model the variability in the F0 parameter values of emotional utterances. Using emotional regions one can theoretically represent how the F0 contour of an utterance can be modified without significantly affecting its emotion and speech quality. Note that, the dimension of these regions is dependent on the number of parameters that are used. In this paper, for easy visualization we worked with two-dimensional (2D) regions, which were estimated based on the F0 mean and range values. If F0 contour shape was also considered as a factor, the emotional regions would be three dimensional.

Grouping the utterances into 16 groups based on speaker, sentence, and emotion, as explained above, for each group, the group mean vector and covariance matrix were calculated and constant Mahalanobis distance contours—equal to 3—were determined. The center and shape of these contours are determined by the mean vectors and by the covariance matrices, respectively. The contours are ellipses (Fig. 3) and they represent the equal probability density Gaussians (Duda *et al.*, 2001). The Mahalanobis distance was set to 3 as a result of experiments that showed that these contours were reliable estimates for the distribution of resynthesized utterances as can be seen in Fig. 4.

Each of these Gaussian emotional regions, shown in Fig. 3, represent a subset of possible F0 values with which a given original utterance can be modified to maintain the same emotion perception by the majority of the listeners. Note that the Gaussian emotional regions in Fig. 3 are considered a subset of the true emotional regions because they were estimated based on a limited set of modifications (listed in Table I).

Speech quality can be also included as one of the factors determining the emotional regions. In this case, in addition to the requirement that the utterances need to be perceived as conveying a certain emotion, they are also required to be perceived with a certain minimum average speech quality.

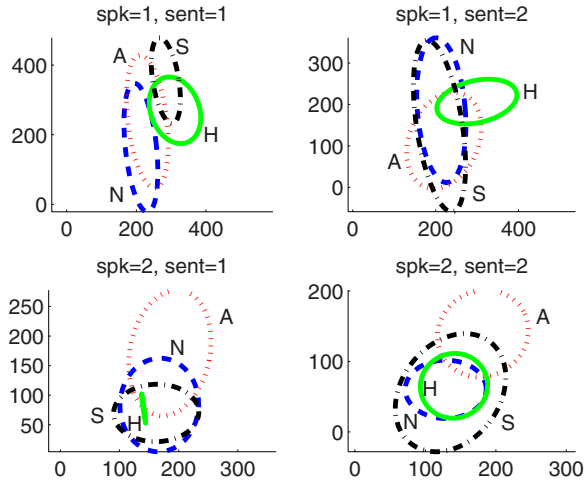

FIG. 3. (Color online) The Gaussian emotional regions for each emotion, speaker, and sentence. x axis=F0 mean (Hz), y axis=F0 range (Hz).

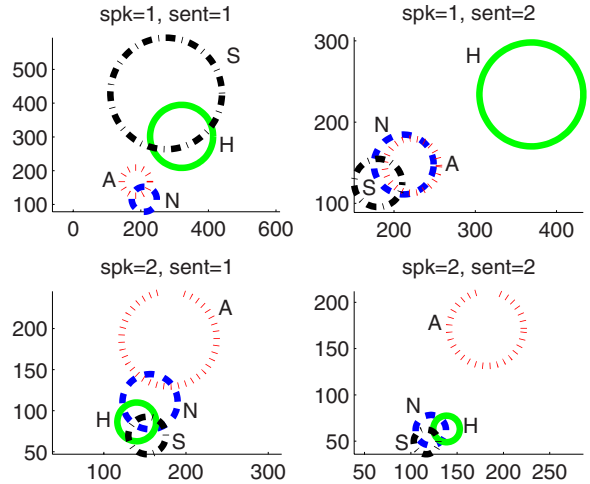

FIG. 5. (Color online) Euclidean emotional regions estimated from the F0 contours of original utterances. x axis=F0 mean (Hz), y axis=F0 range (Hz).

For example, denoting the average speech quality by ρ, it may be required for each of them to satisfy $\rho \geq 3$, $\rho \geq 3.5$, $\rho \geq 4$, or $\rho \geq 4.5$ conditions. Under these requirements, the area of emotional regions can be expected to decrease as quality requirements increase. An example is shown in Fig. 4, which shows the emotional regions for angry utterances. Although not shown, when higher quality conditions were applied, the size of the emotional regions (shown in Fig. 3) decreased in a similar manner for the other emotions as well.

The emotional regions shown in Fig. 3 and Fig. 4 were estimated using the resynthesized utterances. In order for them to be used in real life applications they need to be estimated automatically for individual utterance F0 contours. For that purpose, F0 mean, range, and std values can be used. For a given utterance, representing the center by [F0 mean, F0 range], and the radius by F0 std, (circular) Euclidean emotional regions can be constructed. As shown in Fig. 5, contours of one std Euclidean distance from the [F0 mean, F0 range] point were plotted and considered as Euclidean emotional regions for a given utterance.

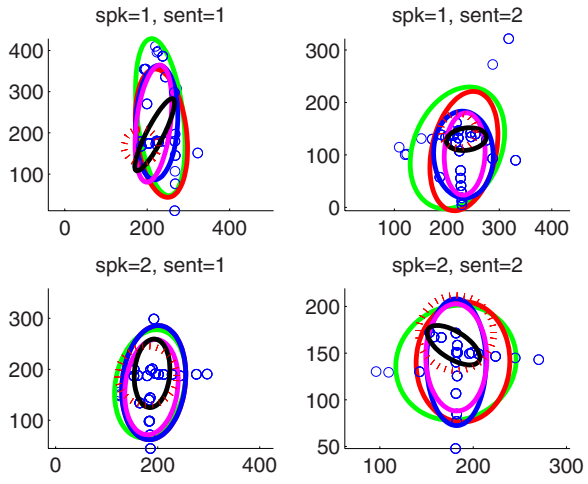

FIG. 4. (Color online) Emotional regions for different speech quality (ρ) requirements for angry emotion. The areas of emotional regions decrease as quality requirements increase. Displayed are the following quality conditions: (1) No restriction (same as Fig. 3), (2) $\rho \geq 3$, (3) $\rho \geq 3.5$, (4) $\rho \geq 4$, (5) $\rho \geq 4.5$. Small circles show the resynthesized utterances. x axis=F0 mean (Hz), y axis=F0 range (Hz).

In order to determine how well the Euclidean regions can approximate the Gaussian regions, they were plotted together in Fig. 6. Comparing the two regions we note that for most of the groups Euclidean regions lie inside the Gaussian regions. This shows that the regions estimated by the Euclidean method are reasonable and accurate, representative of a subset of Gaussian regions. This is more clearly seen in Fig. 4 where the Gaussian regions for different quality conditions were plotted together with the Euclidean regions (shown as dotted circles). While observing the plots, note the similarity between the Euclidean emotional regions and higher quality ($\rho \geq 4.5$) Gaussian regions. Figure 4 shows the results for angry utterances only, but the results were similar for other analyzed emotions.

From the figures it can be seen that the emotional regions were different for different speakers. For example in Fig. 3, note that for speaker 1, the happy region did not lie inside sad or neutral region, while for speaker 2 it did. Also note that for speaker 2, the intersection between angry and neutral regions was smaller in comparison to their intersection for speaker 1. In addition to the speaker related differences, differences due to sentences were also observed. For instance, for speaker 1, the neutral region for sentence 2 was inside the sad region, while for sentence 1 it was not.

The differences between the emotional regions can be attributed to the differences in the factors—such as sentence, speaker, and emotion—that affect the F0 contour characteristics (see Fig. 1). The effects of these factors are examined in detail in the next section (Sec. V) where statistical analyses results are presented.

It is important to note that the present emotional regions are proposed as models to represent the variability of single utterance F0 parameter values, and therefore they are specific to the utterance itself. They show the ranges within which the utterance F0 parameters can be modified without affecting its emotional and speech quality. However, they do not

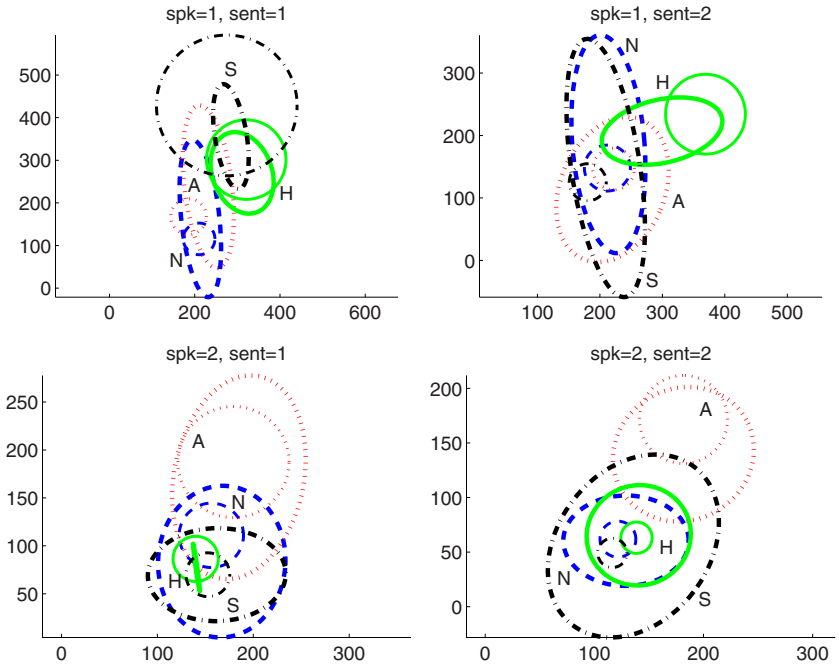

FIG. 6. (Color online) The perceived Gaussian emotional regions and estimated Euclidean emotional regions, x axis=F0 mean (Hz), y axis=F0 range (Hz).

necessarily show how these parameters should be modified to synthesize speech with a new emotion. For example, if a happy utterance is modified so that its new F0 values fall outside the happy emotional region, it is known that it will not be perceived as happy anymore. However, it is not necessarily true that if the new point is in the neutral (or any other) region then the utterance will be perceived as neutral. This is due to the fact that the perception of emotions is based not only on F0, but on the combined effects of prosodic, spectral, and linguistic factors. Therefore only when all of these factors are used to construct multidimensional regions one can predict the emotion only from the region itself. Note that the present emotional regions can be considered as projections of the hypothetical multidimensional regions on the linguistic (sentence), spectral (speaker), and F0 planes.

As shown in Fig. 4, in general, the Euclidean emotional regions can be considered to be reasonable approximations to the high quality Gaussian regions. As seen in Fig. 5, in the Euclidean method the assumption is that the variations in F0 mean and F0 range directions are equal. If needed, another model which estimates possible F0 range and F0 mean variations separately can be also constructed. For example, for an utterance, if in addition to the F0 contour, word (and/or syllable) boundaries are also known, one can calculate the F0 mean and range values for each word (syllable). Then, these two vectors (the vectors of F0 means and F0 ranges) can be used to calculate the covariance matrix, which can be used to form the new emotional regions (which will be ellipsoidal and not circular) for the utterance.

V. STATISTICAL ANALYSIS OF EMOTION AND SPEECH QUALITY PERCEPTION

In order to examine the effects of utterance emotion, speaker, sentence, and modification factors on emotion and quality perception a four-way ($4\times2\times2\times30$) repeated measures analysis of variance (ANOVA) model was designed. The model consisted of four independent variables [*original utterance emotion (4)*, *speaker (2)*, *sentence (2)*, and *modification (30)*] used as repeated measures, and two dependent variables (*emotion* and *speech quality*). The model was fully counterbalanced across seven subjects (i.e., listeners) that evaluated all 480 utterances (which correspond to all possible combinations of the independent variables).

Of the within-subject independent variables, *utterance emotion type* has four levels that reflect the emotions intended by the speakers. These levels correspond to happy, angry, sad, and neutral emotions. The *speaker* variable has two levels, corresponding to speaker 1 (who was a female) and speaker 2 (who was a male). The *sentence* variable has two levels, sentence 1 (She told me what you did.), and sentence 2 (This hat makes me look like an aardvark.). And finally, the *modification* variable has 30 levels that correspond to all performed F0 modifications cases (29), plus the no modification (i.e., original) case (see Table I for the complete list of the modifications).

There are two dependent variables. Of these, *emotion selection* is a nominal variable that was defined as a dichotomous outcome reflecting whether the emotion selected by a listener for a resynthesized utterance was the same as the emotion of the original utterance. If they were the same the variable was set to 1, if they were different it was set to 0. A dichotomous variable was used, because the purpose of the experiment was to investigate specifically the role of the F0 component in the perception of the original emotion.

The *quality* dependent variable was used as a measure of the perceived speech quality that was evaluated on a five point scale as explained in Sec. III.

TABLE II. Cochran's Q statistics calculated for *emotion selection* dependent variable. Significant results are in *italic* form.

Modification		Mean		Range		Stylization	
Emotion	Spk./Sent.	Sent1	Sent2	Sent1	Sent2	Sent1	Sent2
Happy	spk1	Q(14)=16.26 p=0.298	*Q(14)=27.06 p=0.019*	*Q(10)=20.00 p=0.029*	Q(10)=17.61 p=0.062	*Q(5)=19.52 p=0.002*	Q(5)=10.77 p=0.056
	spk2	*Q(14)=29.63 p=0.009*	*Q(14)=32.36 p=0.004*	Q(10)=16.79 p=0.079	Q(10)=9.00 p=0.532	Q(5)=10.91 p=0.053	Q(5)=4.00 p=0.549
Angry	spk1	*Q(14)=28.24 p=0.013*	Q(14)=11.41 p=0.654	*Q(10)=38.31 p<0.001*	Q(10)=10.00 p=0.440	Q(5)=5.00 p=0.416	Q(5)=10 p=0.075
	spk2	Q(14)=14.25 p=0.431	*Q(14)=30.84 p=0.006*	*Q(10)=33.08 p<0.001*	Q(10)=14.70 p=0.144	Q(5)=3.46 p=0.629	*Q(5)=18.10 p=0.003*
Sad	spk1	Q(14)=11.17 p=0.673	*Q(14)=31.31 p=0.005*	Q(10)=7.78 p=0.651	Q(10)=4.24 p=0.936	Q(5)=3.40 p=0.639	Q(5)=5.56 p=0.352
	spk2	*Q(14)=32.62 p=0.003*	Q(14)=16.63 p=0.277	Q(10)=15.22 p=0.124	Q(10)=7.14 p=0.712	Q(5)=8.23 p=0.144	*Q(5)=15.00 p=0.010*
Neutral	spk1	*Q(14)=29.38 p=0.009*	*Q(14)=30.92 p=0.006*	*Q(10)=24.00 p=0.008*	*Q(10)=29.34 p=0.001*	*Q(5)=15.85 p=0.007*	*Q(5)=21.07 p=0.001*
	spk2	Q(14)=21.33 p=0.093	*Q(14)=26.80 p=0.020*	Q(10)=11.72 p=0.304	Q(10)=10.00 p=0.440	*Q(5)=15.29 p=0.009*	Q(5)=6.30 p=0.278

A. Factors influencing emotion perception

The null hypothesis tested was the following: *The probability of intended (i.e., original) emotions correctly perceived by listeners is equal across different variants in a group*. The variants in this case, as explained above, consisted of all possible combinations of independent variables. There were 480 different variants in total, which were grouped based on the sentence, speaker, emotion, and modification (mean, range, or stylization) factors, resulting in 48 ($=2\times2\times4\times3$) groups.

In our experimental setup each of the seven listeners (i.e., subjects) evaluated all of the utterances. Thus, the subjects were treated as related samples. Therefore, to test the null hypothesis, Cochran's Q test was used. The required condition for the application of Cochran's Q test, that the number of the conditions (K) and number of the listeners (N) are such that $KN>30$, was satisfied for all of the analyzed groups. The results of these tests are shown in Table II. The statistically significant ($p<0.05$) results are shown in *italics* for ease of differentiation.

Note that, since the purpose of this analysis was to investigate whether F0 modifications are sufficient to alter the emotional content of original utterances, for each of the cases compared in Table II, the original (i.e., unmodified) utterances were also included. For example, the results reported in the lower right corner (of Table II) are for the group consisting of neutral sentence 2 recorded by speaker 2 and its stylization modifications, in total 6 utterances (5 modified and one original). The size of the groups comparing mean, range, and stylization modifications were 15, 11, and 6, respectively.

From the results it is observed that the effects of F0 modifications on emotion perception were dependent on *emotion*, *speaker*, and *sentence* factors, showing the complex interactions between these parameters. For example, it is seen that sentence 2 uttered in angry emotion was not significantly influenced by the range modifications, while in contrast, the same modifications caused the perceived emotions for angry sentence 1 to be significantly different than its original. Note, however, that when sentence 2 uttered by speaker 1 in neutral style was modified by the same F0 range modifications, the perceived emotions were different than the emotions perceived for the original utterance. Other similar observations can be made from the results in Table II.

Also it is notable that especially for speaker 1 modifying the F0 characteristics of neutral utterances caused the perception of new emotional nuances. In contrast, this result was less common for angry and happy emotions, and the least common for sad emotion.

B. Factors influencing speech quality perception

The null hypothesis tested was the following: *The mean of the perceived quality is the same under different conditions*. The repeated measures ANOVA results are reported in Table III (the significant results are shown in *italic*). Shown in the tables are the F values calculated from Greenhouse–Geisser tests. This test was preferred because it accounts—by adjusting the degrees of freedom—for the violations of sphericity condition.

The results show that the main effects of emotion, speaker, and sentence factors were insignificant, while the main effect of modification was significant (see Table III). Interesting results were found from the interaction analysis of the within-subject factors. Note, for instance, that the effect of F0 modifications (on the perceived speech quality) was significantly dependent on emotion, speaker, and sentence variables. Also note that the effect of speakers was significantly dependent on emotion, but not on sentence.

In order to analyze the effects of F0 modifications, speaker and sentence factors for different emotion conditions, statistical analyses were performed separately for different emotions. These results are shown in Table IV. For all emotions, it is seen that the main effect of speaker was not statistically significant. In contrast, the main effect of modi-

TABLE III. Repeated measures ANOVA statistics calculated for *quality* dependent variable. The reported are the F values for Greenhouse–Geisser tests.

Factor	Greenhouse–Geisser statistics
Emotion	$F(1.72,10.30)=1.87$, $p=0.203$
Speaker	$F(1,6)=0.96$, $p=0.366$
Sentence	$F(1,6)=0.02$, $p=0.890$
Modification	$F(3.84,23.02)=28.27$, $p<0.001$
*Emo*Spk*	$F(2.50,15.03)=3.64$, $p=0.043$
*Emo*Sent*	$F(1.20,7.21)=7.70$, $p=0.024$
*Emo*Modif*	$F(5.40,32.40)=6.07$, $p<0.001$
Spk*Sent	$F(1,6)=3.144$, $p=0.127$
*Spk*Modif*	$F(4.24,25.45)=13.25$, $p<0.001$
*Sent*Modif*	$F(4.12,24.74)=5.16$, $p=0.003$
*Emo*Spk*Sent*	$F(1.22,7.30)=7.21$, $p=0.027$
*Emo*Spk*Modif*	$F(5.43,32.60)=2.64$, $P=0.037$
*Emo*Sent*Modif*	$F(5.20,31.22)=2.76$, $p=0.034$
*Spk*Sent*Modif*	$F(4.67,28.06)=3.36$, $p=0.019$
*Emo*Spk*Sent*Modif*	$F(4.89,29.35)=2.84$, $P=0.034$

fication was significant in all cases. Interestingly, we also observe that the effect of sentence was significant for angry and neutral emotions, but not for happy and sad. In fact, note that the patterns of significant results were the same for happy and sad emotions and somewhat similar between angry and neutral emotions.

VI. EFFECTS OF F0 MODIFICATIONS ON EMOTIONAL CONTENT

In this section, the effects of F0 modifications are compared in terms of emotional content that was perceived. Responses from all 14 listeners were included in these evaluations.

In Fig. 7 the changes in the emotion recognition percentages observed after each modification are shown. The change was defined as the difference between recognition percentages of unmodified and modified utterances. Chi-square tests with 95% confidence interval were used to calculate whether or not the change was significant. The discussions below focus mainly on the significant modifications.

The mean modifications that caused significant ($p<0.05$) emotion perception changes for speaker 1 were m4, m8, m9, m10, m11 [Fig. 7(a)]. These results show that speaker 1 was quite robust against the F0 mean modifications. It was only when the F0 mean was changed by $\pm 50\%$ significant changes were observed. Increasing F0 mean caused the *neutral* and *angry* recognition percentages to drop, and *sad* and *other* recognition percentages to increase. Interestingly, adjusting the mean to be in 50–150 Hz range caused increase in *happy* and *other* responses. Note that in all these instances the speech quality degraded significantly [Fig. 7(e)].

For speaker 2—as seen with speaker 1—increasing or decreasing F0 mean by 50% caused an increase in *sad* and *other* perception percentages [Fig. 7(c)]. It was also observed that some of the modifications caused an increase in the *neutral* and *other* responses, but not in the *happy* or *angry* responses. The statistically significant modifications in this case were m1, m4, m7, m8, m9, m10, m13, and m14. All of

TABLE IV. Repeated ANOVA statistics calculated for quality dependent variable. The reported are the F values for Greenhouse–Geisser tests. Significant results are shown in *italic* for easy differentiation.

Factor	Happy	Angry	Sad	Neutral
Speaker	$F(1,6)=1.66$	$F(1,6)=0.75$	$F(1,6)=2.13$	$F(1,6)=0.03$
	$p=0.245$	$p=0.421$	$p=0.195$	$p=0.864$
Sentence	$F(1,6)=0.85$	$F(1,6)=11.97$	$F(1,6)=0.55$	$F(1,6)=21.26$
	$p=0.392$	$p=0.013$	$p=0.486$	$p=0.004$
Modification	$F(3.68,22.06)=16.84$	$F(4.23,25.38)=26.93$	$F(4.63,27.79)=7.97$	$F(4.52,27,12)=24.5$
	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
Spk*Sent	$F(1,6)=1.92$	$F(1,6)=33.51$	$F(1,6)=3.67$	$F(1,6)=9.72$
	$p=0.215$	$p=0.001$	$p=0.104$	$p=0.021$
Spk*Modif	$F(4.43,26.56)=6.98$	$F(4.21,25.27)=8.46$	$F(4.72,28.33)=3.05$	$F(5.01,30.07)=7.59$
	$p<0.001$	$p<0.001$	$p=0.027$	$p<0.001$
Sent*Modif	$F(4.75,28.50)=2.12$	$F(4.69,28.12)=8.36$	$F(4.72,28.29)=1.87$	$F(5.02,30.13)=2.35$
	$p=0.095$	$p<0.001$	$p=0.135$	$p=0.065$
Spk*Sent*Modif	$F(4.14,24.82)=1.55$	$F(4.41,26.44)=2.64$	$F(4.51,27.03)=2.57$	$F(4.74,28.46)=5.23$
	$p=0.217$	$p=0.051$	$p=0.055$	$p=0.002$

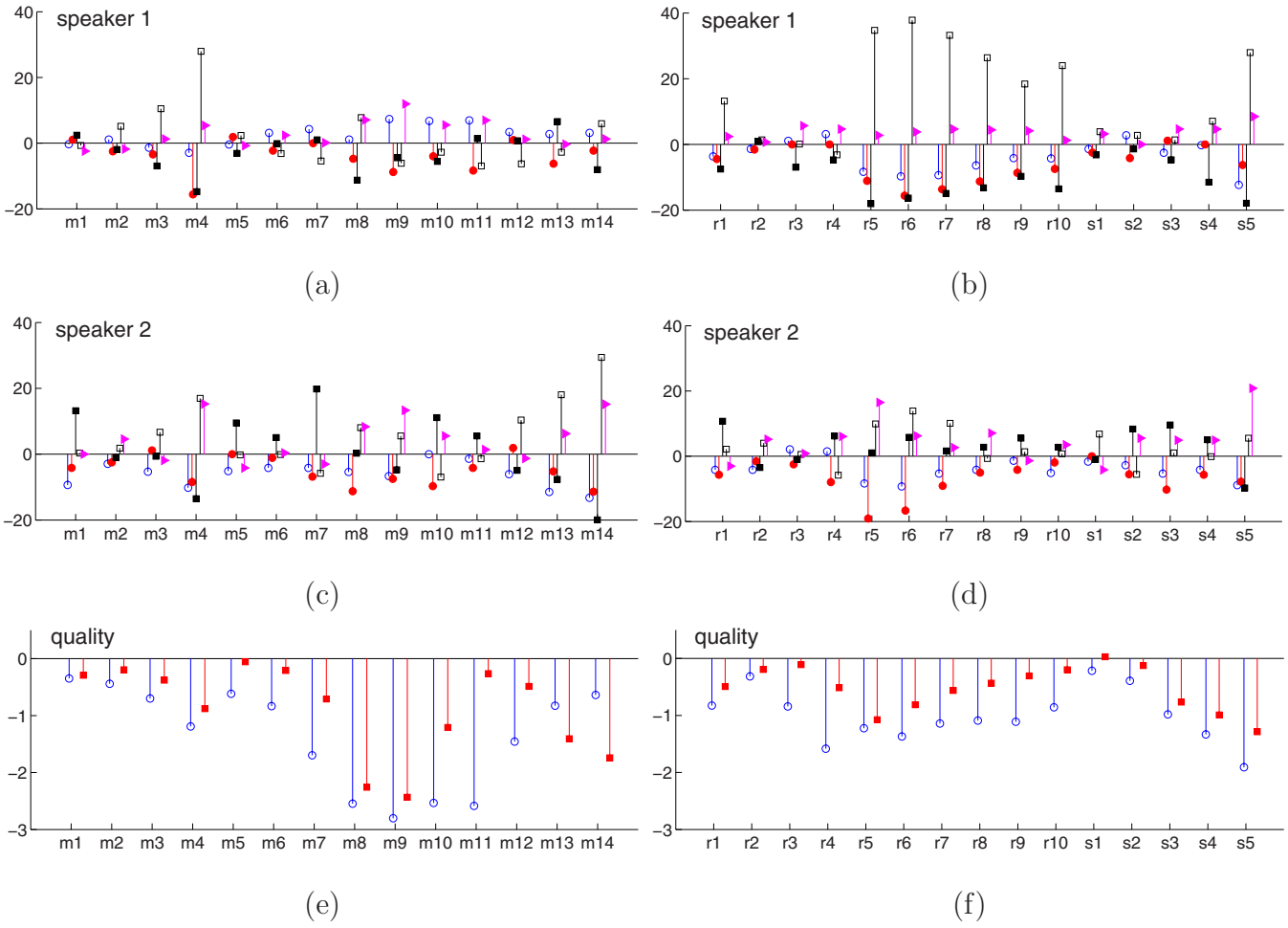

FIG. 7. (Color online) Figures (a), (b), (c), (d): The differences between the emotion recognition percentages of original and modified utterances. Happy=open circle, Angry=filled circle, Sad=open square, Neutral=filled square, Other=filled triangle. Figures (e), (f): The differences between the average speech qualities (5=excellent, 4=good, 3=fair, 2=poor, 1=bad) of original and modified utterances. Speaker 1=open circle, Speaker 2=filled square.

these (except m1) caused a significant drop in the speech quality.

The effects of F0 range modifications were more prominent than F0 mean. For speaker 1, decreasing the F0 range by more than 50% caused a significant increase in *sad* responses (r1, r5, r6, r7, r8, r9, r10). The effect of the F0 range on the sad emotion percentages was continuous (Ladd *et al.*, 1985) and it could be easily parametrized [Fig. 7(b)]. The drop in the perceived speech quality was less severe than F0 mean modifications, suggesting that one should perform range and not mean modifications during the synthesis of emotional speech.

The effects of range modifications on speaker 2 were also significant, however not as strong as they were for speaker 1 [Fig. 7(d)]. This can be attributed to the lower F0 range of this speaker. The modifications that caused significant emotion perception difference were r4, r5, r6, r7, r8. These modifications increased *sad* perception, and decreased *happy* and *angry* perception. In contrast to speaker 1, some of them (r1, r6) also increased the *neutral* perception.

Interesting results were observed for stylization modifications. For speaker 1 [Fig. 7(b)], only s4 and s5 caused significantly different results. An increase in the *sad* and *other* responses and drop in quality was seen for these cases. These results show that eliminating the small prosodic variations (s1, s2, s3) in the F0 contour shape did not significantly decrease the perception of the original emotions. It was only when the F0 contour at the sentence level was fully linearized (as seen in Fig. 2)—eliminating any accents and foot patterns (Klabbers and van Santen, 2004)—the percentages of *happy* and *angry* emotions started to decrease. In these cases the utterances were mostly perceived as *sad* or *other*.

This is an important result which has implications for emotional speech synthesis. As shown in our previous work (Bulut *et al.*, 2005, 2002), for synthesis of emotions such as anger and happiness, in addition to prosody, spectral characteristics also play an important role. Therefore, during synthesis of these emotions one needs to concentrate more on the overall F0 contour shape, F0 range, and spectral characteristics and can ignore the small prosodic variations in the F0 contour shape. As we show later, these small prosodic variations were more important for high quality perception than emotion perception.

The arguments above were also valid for the speaker 2, for whom only some particular stylization modifications (s2, s3, s5) caused significant changes [Fig. 7(d)], with minimal degradation in quality [Fig. 7(f)]. In these cases increase in the *other* responses was accompanied either by increase in

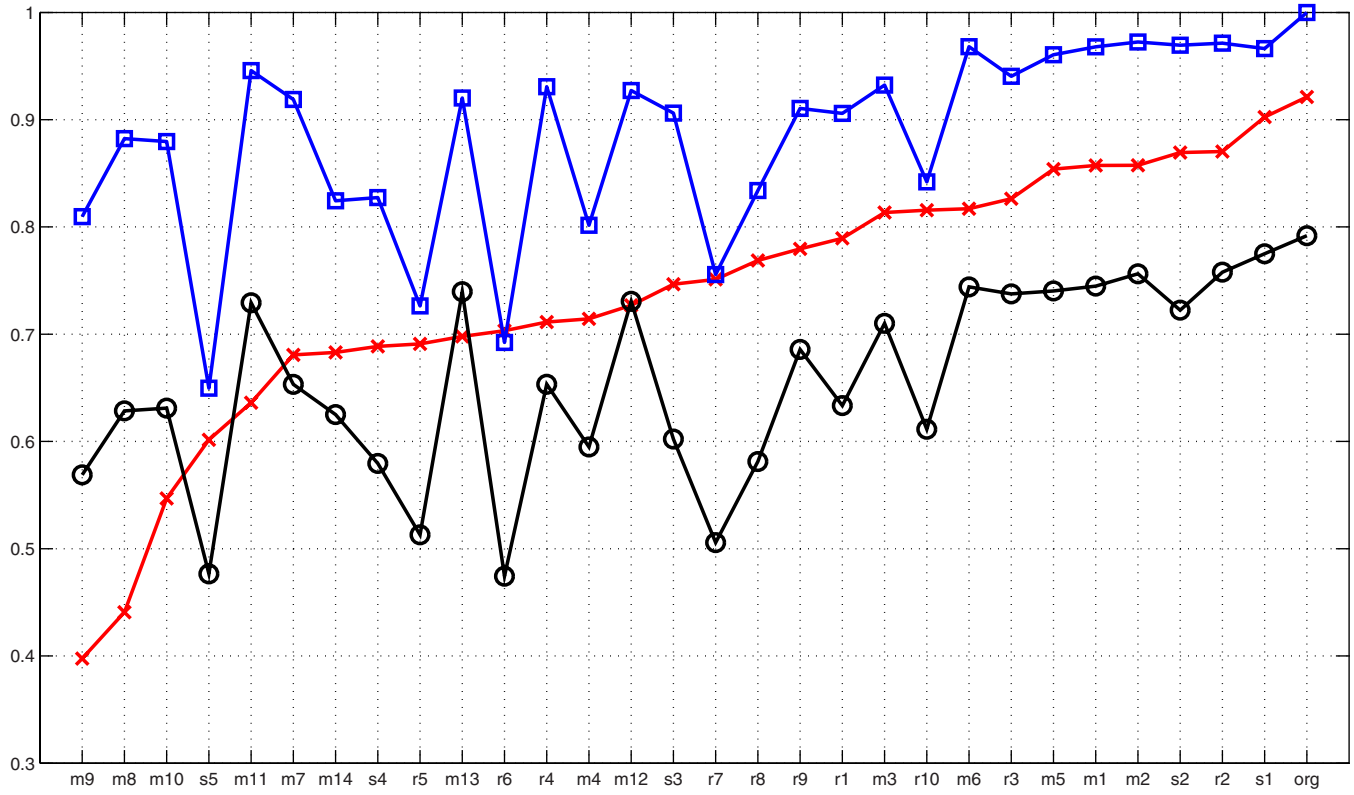

FIG. 8. (Color online) Relation between *average quality*, *similarity*, and *percentage* parameters. Note that the quality is normalized: 1=Excellent, 0.8=Good, 0.6=Fair. 0.4=Poor. 0.2=Bad. Squares (□) are used for *similarity*, circles (○) for *percentage*, and (*x*) for *quality* variables.

sad or *neutral* responses. It was particularly interesting to note that both s1 and s4, which caused significantly different F0 contour shapes, did not cause any significant emotion changes. That the s4 modification was not significant, while s3 was, was unexpected and it can be attributed to the variability inherent in the subjective nature of the listening tests.

As expected, the modifications that received high quality responses were the ones that did not cause any significant changes in the emotional content. Significant emotion change was in general accompanied with the degradation in quality. However in some instances, especially when the F0 range of speaker 2 was modified, despite significant emotion change the quality was not affected.

In almost all instances, the original emotions were correctly recognized by the majority (i.e., 50% or more) of the listeners. This shows that despite the quality distortions the original emotions were still well perceived.

In order to visualize these relations between quality and emotion perception, we define two new variables, *percentage* and *similarity* (see Fig. 8). The percentage variable represents the percentage of listeners that perceived the same emotion as the original emotion. The similarity variable is a measure that is the cosine of the angle between two vectors and has a large value (i.e., close to 1) when the vectors point in the same direction (Duda *et al.*, 2001). It is calculated using Eq. (1), where x and y are vectors, of size $[5\times1]$, showing the fractions of perceived emotions, for an original (x) and a modified (y) utterance, respectively. For example, a vector $y=[0.5\ 0.3\ 0.1\ 0\ 0.1]$ was used for an utterance that was perceived as happy, angry, sad, neutral, and other by 50%, 30%, 10%, 0%, and 10% of human raters, respectively.

$$s(x,y)=\frac{x^t y}{\|x\|\|y\|}. \tag{1}$$

In summary, the effects of F0 range modifications were more significant than F0 mean modifications. Stylization modifications were also effective, but only when performed in large semitone scales. They showed that small prosodic variations in the F0 contour shape were more related to the quality of speech and not to its emotional content.

VII. DISCUSSION

In speech synthesis, it is important to study the role of the acoustic parameters in connection with the human perception of prosodic and paralinguistic features (Picard, 1997; Picard *et al.*, 2004; Roach, 2000; Traunmuller, 2005). The results of this paper show that in order to be able to describe F0 variations occurring in emotional speech *sentence, speaker*, and *emotion* factors should be considered. These are the factors that determine how emotional regions (Sec. IV) will be shaped.

Sentence (i.e., linguistic content) should be taken into account because—together with speaker and emotion characteristics—sentence structure [i.e., focus, modality, length (Pell, 2001)] determines how the pitch (and also duration, energy, formant frequencies, and meaning) will be generated.

Instead of including the linguistic content as a factor in the analysis, an alternative approach is to minimize its ef-

fects. One way to do that is by using nonsense sentences (Banziger and Scherer, 2006). This method eliminates the semantic effects, however it also may cause the acoustic parameters (e.g., F0, duration, energy) to be modulated in an unnatural fashion. Therefore, the results may not be easily generalizable to real life utterances.

Probably a better parametrization of emotions can be achieved not by restricting the variance in the different features but by restricting the emotion space itself. This may be achieved by defining more homogeneous emotion categories. One good example is given by Banziger and Scherer (2006), who, in addition to the classic categorical emotion labels, also used activation level differences (Grimm *et al.*, 2007) to describe the emotions. This suggests that in order to better relate the acoustic parameter variation to particular emotions, a hybrid labeling scheme combining categorical (Ekman and Friesen, 1977) and attribute descriptions (Schlosberg, 1954) can be utilized. For example, considering the findings showing that valence, activation, and intensity dimensions are correlated with the acoustic features of emotional speech (Grimm *et al.*, 2007; Schroder *et al.*, 2001), an angry utterance can be described as *angry*, *high (low, medium) activation*, *high (low, medium) valence*, *high (low, medium) intensity*, instead of just *angry*.

Having a better description for emotions can be expected to produce smaller emotional regions. Smaller regions can be expected to overlap less, which in turn will help to better parameterize and differentiate between different emotions in terms of their acoustic features. For example, evaluating angry speech as high or low activation anger would have created two emotional regions instead of one, which theoretically would have helped to better describe how F0 characteristics relate to the angry emotional content. As shown in this paper, the significant overlap between the regions of emotions labeled using the categorical labels indicates that a hybrid labeling technique is necessary for future research in this area.

Considering the small number of sentences and speakers that were analyzed in this study, our future plan is to perform similar studies on a larger dataset. Also, we plan to perform similar analyses for the duration and energy parameters. Increasing the number of sentences, speakers, emotions, and acoustic parameters will provide better information about how the interactions between different factors can be described and parametrized.

VIII. CONCLUSION

The variability of pitch (and therefore F0 contour) is one of the basic characteristics of natural human speech. It has been shown that the same text recorded at different times can have very different F0 characteristics. In this study using an analysis by synthesis method we showed the variability that exists in the F0 mean, range, and shape parameters of emotional speech. The results showed that even significant variation in F0 parameters did not mask the original emotion perception. It was observed that F0 modification caused *sad*, *neutral* or *other* emotion perception to increase, and *angry* or *happy* perception to decrease. The effects of F0 range modifications on emotion perception were more prominent than F0 mean modifications. Also, for F0 range modifications, the drop in the perceived speech quality was less than F0 mean modifications. These results suggest that one should focus on range and not mean modifications during the synthesis of emotional speech. The results were significantly dependent on the speaker and the original utterance characteristics.

In order to model the observed variability in the F0 contour, an emotional regions approach was introduced. The observed emotional regions derived from the data were represented as 2D Gaussian ellipses which showed the limits within which the F0 contour of a given utterance can be modified. In order to model these observed regions, Euclidean emotion regions based on F0 statistics (mean, range, std) were proposed. It was shown that the Euclidean regions can be used as reliable approximations to the high quality Gaussian emotional regions.

The emotional regions concept can be applied to the other acoustic parameters as well. If duration and spectral envelope variations are modeled together with energy and F0 variations, it will be possible to build multidimensional emotional regions for each emotion, which can then be used in emotion conversion and synthesis of speech. This is a task for our future research.

Banziger, T., and Scherer, K. R. (**2006**). "The role of intonation in emotional expressions," Speech Commun. **46**, 252–267.

Boersma, P., and Weenink, D. (**2007**). "Praat: doing phonetics by computer (version 4.5.18) [computer program]," URL http://www.fon.hum.uva.nl/praat/, last retrieved March 9, 2008.

Braun, B., Kochanski, G., Grabe, E., and Rosner, B. S. (**2006**). "Evidence for attractors in English intonation," J. Acoust. Soc. Am. **119**(6), 4006–4015.

Bulut, M., Busso, C., Yildirim, S., Kazemzadeh, A., Lee, C. M., Lee, S., and Narayanan, S. (**2005**). "Investigating the role of phoneme-level modifications in emotional speech resynthesis," in *Proc. of Eurospeech, Interspeech*, Lisbon, Portugal.

Bulut, M., Narayanan, S., and Syrdal, A. K. (**2002**). "Expressive speech synthesis using a concatenative synthesizer," in *International Conference on Spoken Language Processing*, Denver.

Burkhardt, F., and Sendlmeier, W. F. (**2000**). "Verification of acoustical correlates of emotional speech using formant-synthesis," in *ISCA Workshop on Speech and Emotion*, New Castle, Northern Ireland, UK.

Cahn, J. E. (**1990**). "The generation of affect in synthesized speech," J. Am. Voice I/O Soci. **8**, 1–19.

Chu, M., Zhao, Y., and Chang, E. (**2006**). "Modeling stylized invariance and local variability of prosody in text-to-speech synthesis," Speech Commun. **48**, 716–726.

Cowie, R., Cowie, E. D., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. (**2001**). "Emotion recognition in human-computer interaction," IEEE Signal Process. Mag. **18**(1), 32–80.

Davitz, J. R. (**1964**). *The Communication of Emotional Meaning* (MCGraw–Hill, New York).

Duda, R. O., Hart, P. E., and Stork, D. G. (**2001**). *Pattern Classification*, 2nd ed. (Wiley-Interscience, New York).

Ekman, P., and Friesen, W. V. (**1977**). *Manual for the Facial Action Coding System* (Consulting Psychologist Press, Palo Alto).

Grimm, M., Mower, E., Kroschel, K., and Narayanan, S. (**2007**). "Primitives based estimation and evaluation of emotions in speech," Speech Commun. **49**, 787–800.

Iida, A., Campbell, N., Higuchi, F., and Yasumura, M. (**2003**). "A corpus-based speech synthesis system with emotion," Speech Commun. **40**, 161–187.

Klabbers, E., and van Santen, J. P. H. (**2004**). "Clustering of foot-based pitch contours in expressive speech," in *Proc. of the 5th ISCA Speech Synthesis Workshop*, Pittsburg.

Ladd, D. R., Silverman, K. E. A., Tolkmitt, F., Bergmann, G., and Scherer, K. R. (**1985**). "Evidence for the independent function of intonation con-

tour type, voice quality, and F0 range in signaling speaker affect," J. Acoust. Soc. Am. **78**(2) 435–444.
Montero, J. M., Gutierrez-Arriola, J., Colas, J., Enriquez, E., and Pardo, J. M. (**1999**). "Analysis and modeling of emotional speech in Spanish," in *International Congress of Phonetic Sciences*, San Francisco.
Moulines, E., and Charpentier, F. (**1990**). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun. **9**, 453–467.
Murray, I. R., and Arnott, J. L. (**1993**). "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," J. Acoust. Soc. Am. **93** 1097–1108.
Pell, M. D. (**2001**). "Influence of emotion and focus location prosody in matched statements and questions," J. Acoust. Soc. Am. **109**(4), 1668–1680.
Picard, R. (**1997**). *Affective Computing* (MIT Press, Cambridge, MA).
Picard, R. W., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D., Machover, T., Resnick, M., Roy, D., and Strohecker, C. (**2004**). "Affective learning – a manifesto," BT Technol. J. **22**(4), 253–269.
Raux, A., and Black, A. (**2003**). "A unit selection approach to F0 modeling and its application to emphasis," in *Proc. of ASRU* (St. Thomas, U.S. Virgin Islands).
Roach, P. (**2000**). "Techniques for the phonetic description of emotional speech," in *ISCA Workshop on Speech and Emotion*, Newcastle. Northern Ireland, UK.
Scherer, K. R. (**2003**). "Vocal communication of emotion: A review of research paradigms," Speech Commun. **40**(1–2), 227–256.
Schlosberg, H. (**1954**). "Three dimensions of emotion," Psychol. Rev. **61**, 81–88.
Schroder, M. (**2001**). "Emotional speech synthesis - a review," in *Eurospeech* (Aalborg, Denmark).
Schroder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., and Gielen, S. (**2001**). "Acoustic correlates of emotion dimensions in view of speech synthesis," in *Eurospeech* (Aalborg, Denmark).
Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierre-humbert, J., and Hirschberg, J. (**1992**). "ToBI: A standard for labeling English prosody," in *International Conference on Spoken Language Processing*, Banff, Alberta, Canada, pp. 867–870.
Taylor, P. (**2000**). "Analysis and synthesis of intonation using the tilt model," J. Acoust. Soc. Am. **107**, 1697–1714.
Traunmuller, H. (**2005**). "Speech considered as modulated voice," URL http://www.ling.su.se/STAFF/hartmut/aktupub.htm, revised manuscript (last retrieved **March 9, 2008**).
Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., and Narayanan, S. (**2004**). "An acoustic study of emotions expressed in speech," in *International Conference on Spoken Language Processing*, Jeju, Korea.