

Analysis of Affective Speech Recordings using the Superpositional Intonation Model

Esther Klabbers, Taniya Mishra, Jan van Santen

Center for Spoken Language Understanding
OGI School of Science & Engineering at OHSU
20000 NW Walker Road, Beaverton, OR, 97006, USA

klabbers@cslu.ogi.edu

Abstract

This paper presents an analysis of affective sentences spoken by a single speaker. The corpus was analyzed in terms of different acoustic and prosodic features, including features derived from the decomposition of pitch contours into phrase and accent curves. It was found that sentences spoken with a sad affect were most easily distinguishable from other affects as they were characterized by a lower F_0 , lower phrase and accent curves, lower overall energy and a higher spectral tilt. Fearful was also relatively easy to distinguish from angry and happy as it exhibited flatter phrase curves and lower accent curves. Angry and happy were more difficult to distinguish from each other, but angry was shown to exhibit a higher spectral tilt and a lower speaking rate. The analysis results provide informative clues for synthesizing affective speech using our proposed recombinant synthesis method.

1. Introduction

Generating meaningful and natural sounding prosody is a central challenge in TTS. In traditional concatenative synthesis, the challenge consists of generating natural sounding target prosodic contours and imposing these contours on recorded speech without causing audible distortions. In unit selection synthesis, the challenge consists of selecting acoustic units from a large speech corpus that optimally match the phonemic and prosodic contexts required. When expanding a prosodic domain from a neutral reading style to more expressive styles, the size of the speech corpus grows exponentially.

We are developing a new approach to speech synthesis, called *recombinant synthesis* (also known as multi-level unit selection synthesis) in which natural prosodic contours and phoneme sequences are recombined using a superpositional framework [13]. The proposed method can use different speech corpora for selecting phoneme units and pitch contour components. As the prosodic space is expanded to include more speaking styles or sentence types (i. e. lists), more pitch contours can be added to the prosodic corpus. The prosodic corpus does not contain the raw pitch contours, as concatenating them would result in audible discontinuities [12], but rather contains phrase curves and accent curves that are derived from the original pitch contour. Recombinant synthesis has advantages over both traditional concatenative synthesis and unit selection in that (i) the pitch contours selected from the database are natural

and smooth, leading to higher quality synthesis, and (ii) much smaller speech corpora are required as the coverage of acoustic and prosodic features is additive instead of multiplicative.

The goal is to select natural-sounding pitch contours that are appropriate for the given context and that are close enough to the original prosody of the selected phoneme units to minimize signal degradation due to pitch modification [5]. This paper discusses preliminary findings related to a set of affective recordings. There have been several studies analyzing affective speech for synthesis purposes [3, 1, 14, 9]. Typically they explore simple prosodic features such as the F_0 mean and range, and phoneme durations. Some studies [9] have gone further and examined pitch contour shapes in different affective conditions. The recordings used in our analysis are by no means complete, nor is the set large enough to make exhaustive predictions, but the analysis method and the acoustic features used to analyze the data will provide valuable information about distinguishing different affects and hopefully will be useful in generating appropriate affective speech. The relevance of acoustic features was analyzed using a repeated measures analysis of variance paradigm and paired t -tests were performed to determine the acoustic differences between pairs of affects.

2. Recordings

This study used a set of affective recordings that was collected for a previous study. A group of 42 actors read 24 sentences in 4 different affects: Angry (A), Happy (H), Fearful (F), and Sad (S). There was considerable variability within subjects with respect to expressing the different affects. For the purposes of speech synthesis of affective speech, one single speaker was chosen for analysis. The chosen speaker is an 8-year old girl who was the most consistent in her renditions of the different affects. This was established in a listening experiment, where 12 people listened to all sentences in random order and assigned affect labels and a confidence score to them.

The speakers did not produce neutral recordings for these 24 sentences. However, the sentences are semantically unbiased in their affective content, i. e., it is impossible to predict which affect is intended from the text alone. Because there are four different versions of each sentence, different affects can be compared side-by-side. The sentences consist of a single phrase 2–5 words in length. The sentences are preceded by short “vignettes” which cue the speaker to produce the correct affect. Table 1 presents 4 example vignettes for one of the sentences. The simulated vocal expressions obtained in this manner will yield more intense, prototypical expressions of affect [14], but for speech synthesis purposes this is desired to ensure correct

This research was conducted with support from NSF grant 0205731, “Prosody Generation in Child-Oriented Speech” and NIH Grant 1R01DC007129, “Expressive and Receptive Prosody in Autism”.

Angry	Happy	Fearful	Sad
The parents had left their teenager home alone for the weekend and had come home to a house that had been turned upside down. The father said angrily:	Her best friend had moved away four months ago. She was contemplating this as the doorbell rang. It was her.	Suddenly the tornado made a turn, and now was heading for where John was standing. 'I'm gonna get killed by a tornado.	She cried when her parents told her that her best friend had been in an automobile accident and may never walk again. She was overcome with grief, and said:
<i>"I don't believe it!"</i>			

Table 1: Affective vignettes for the sentence "I don't believe it".

perceived affects. Moreover, the perception experiment showed that listeners could correctly recognize the intended affects, reflecting the fact that these recordings represent normal expression patterns.

3. Analysis

In this study we used analysis features based on pitch, duration, and energy to distinguish different affects. The pitch values for the recordings were computed using Praat [2]. The advantage of using Praat is that it is able to deal with high frequencies, which are more common in childrens' voices and it allows manual adjustments to the voicing flags on a frame-by-frame basis to obtain the best pitch contour. All resulting pitch contours were manually checked to make sure they were correct. The pitch was used to measure global features such as F_0 mean and range. In addition, more detailed features were computed relating to the phrase curves and accent curves obtained by decomposing the pitch contours according to the superpositional model. The decomposition algorithm will be described in more detail in 3.1.

Phoneme segmentation was performed using CSLU's phonetic alignment system [4]. The phoneme alignment was hand-corrected. The phoneme labeling was used to compute phoneme durations. In addition, the sentences were labeled according to their foot structure. A foot is defined as consisting of an accented syllable followed by all unaccented syllables until the next accented syllable or a phrase boundary. The foot structure could be different in each affect rendition, as the number of accents was not always the same. As a rule, foot labeling was based on the presence of audible emphasis on a syllable. The foot labels were checked by two colleagues to ensure consistency. Phrase-initial unstressed syllables are called *anacrusis*. The accent curves on anacrusis were excluded from our analysis.

Variations in acoustic features between different speaking styles are not restricted to prosody, but also include spectral features such as spectral tilt and spectral balance. Spectral balance represents the amplitude pattern across four different frequency regions. These four bands are generally phoneme independent, and contain the first, second, third and fourth formant for most of the phonemes. Formants contain the largest portion of energy in the frequency domain. Moreover, when some prosodic factors change, e. g., from unstressed to stressed, the energy near formants will be amplified much more than those near other frequency locations. Choosing frequency bands according to formant frequencies has an important advantage for statistical analysis, because it will reduce interactions between phoneme identity and prosodic factors. For speech with 16 kHz sampling rate, the four bands are defined as: B1: 0-800Hz, B2: 800-2500Hz, B3: 2500-3500Hz, B4: 3500-8000Hz. Previous research has

shown systematic variations in spectral balance in phonemes when influenced by syllable stress, word accent, proximity to phrase boundary, and neighboring phonemes [11, 7]. The four band values were computed as an average of three data points nearest to the peak location in the foot. These points were always located in the stressed vowel. The overall energy was computed as a sum of the four bands. The spectral tilt was computed as $\{-2 * B1 - B2 + B3 + 2 * B4\}$. Previous studies have shown that our synthesis system is capable of synthesizing speech with different spectral balance profiles successfully without introducing additional signal degradation [11, 7].

3.1. Decomposition of pitch curves

In the general superpositional model of intonation, the pitch contour is described as the sum of component curves that are associated with different phonological levels, specifically, the phoneme, foot, and phrase level [10, 12]. To apply this model to the recombinant synthesis method, the pitch curves in the prosodic corpus need to be automatically decomposed into their corresponding phrase and accent curves. The phrase curve is the underlying curve that spans an entire phrase. It provides information about the baseline pitch and the global declination. The accent curves span the foot and they convey the amount of emphasis exerted on accented syllables. The typical accent curve template is characterized by an up-down movement in the pitch, although there are also templates for negative accents and phrase-final accents containing continuation rises. Decomposing pitch curves is not trivial, since successive accents may overlap in time and we want to impose as few constraints as possible on the shapes of accent and phrase curves.

The proposed decomposition algorithm has been developed using increasingly more difficult sentences. The first step was to decompose synthetic F_0 contours that were generated with our implementation of the superpositional model and curves generated with the Fujisaki model [12]. The next step was to decompose natural F_0 contours from declarative all-sonorant sentences [8]. The last step involved decomposing natural F_0 contours from unrestricted declarative sentences containing continuation rises [6].

Figure 1 shows the decomposition of the F_0 contours for the sentence "I don't believe it" for all four affects. The estimated F_0 contours, as depicted by the solid continuous lines provide close approximations of the raw pitch contour. The decomposition algorithm optimizes the Root Weighted Mean Square Error (RWMSE) where the weights are determined by the amplitude and voicing flags. The overall RWMSE obtained for this database is 15.65 Hz, which is appropriate given the fact that the recordings are extremely expressive and come from a child whose F_0 excursions occasionally exceeded 800 Hz.

The decomposition takes place on a foot-by-foot basis. The

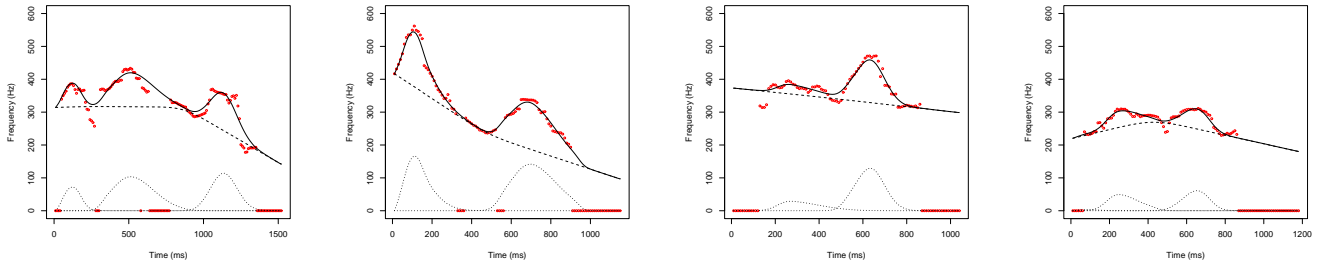


Figure 1: Decomposition of the F_0 contour into a phrase curve and accent curves for the sentence “I don’t believe it”.

Acoustic feature	F -value	p -value	Sig.
Average F_0	15.25	4.06e-08	***
F_0 range	21.85	9.58e-11	***
Phrase curve range	8.39	5.51e-05	***
Average phrase curve slope	5.57	0.0015	**
Start of phrase curve	8.66	4.09e-05	***
End of phrase curve	3.95	0.011	*
Number of accents	1.85	0.14	
First accent amplitude	9.89	1.04e-05	***
Last accent amplitude	9.49	1.63e-05	***
Average accent amplitude	12.68	5.38e-07	***
Speaking rate	1.03	0.38	
Overall energy	29.18	2.62e-13	***
Spectral tilt	7.47	0.00016	***

Table 2: Results for Anova with repeated measures for each acoustic feature. Sig. stands for significance, where * corresponds to a p -value < 0.05 , ** corresponds to a p -value < 0.01 and *** corresponds to a p -value < 0.001

phrase curve consists of piecewise linear segments that are smoothed to create a more natural looking curve. The accent curves are based on generic accent templates which are warped in the time and frequency domain to best match the target curve. Because the sentence content is known and phonemes and feet are labeled, the approximate locations of the accent curves are known. The algorithm requires an approximate location of the accent peak. We obtained initial peak location estimates automatically which were hand-corrected to ensure a close fit.

4. Analysis results

In order to determine which acoustic features were significantly different between affects, an analysis of variance with repeated measures was performed on each acoustic feature. Affect was the dependent variable and sentence number was the error term (because the acoustic features observed are not independent of the sentence content uttered). The analysis of variance results in Table 2 show that most of the features we examined were significantly different across affects. The only features that were not significantly different were the number of accents and the speaking rate. The end value of the phrase curve was only slightly significant.

Most studies on prosody in affective speech ignore the fact that the number of accents might be different across conditions. Informal analysis of the recordings exposed a tendency for speakers to emphasize more words in excited conditions such as angry and happy. Although the number of accents per sen-

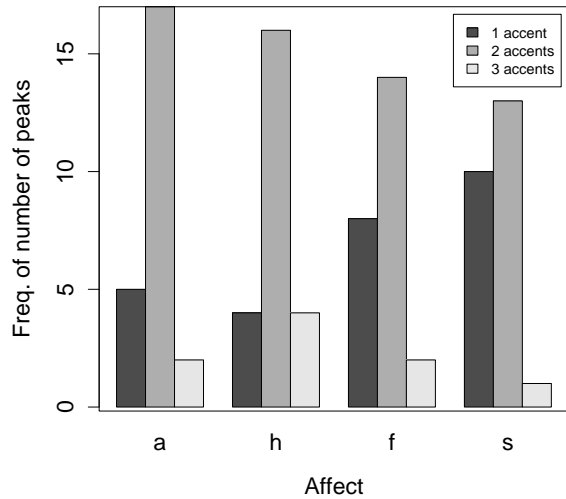


Figure 2: Number of accents per sentence.

tence is not significantly different across affects for the current speaker, there is a clear trend visible in Figure 2. The fearful and sad sentences tend to have fewer accents than the angry and happy conditions. We believe that this trend will become more obvious with longer sentences and text material. The reason it is not significant in this corpus is that the number of stressable words is limited. The analysis of variance presents the overall significance of a feature, but it does not show differences between pairs of affects. Therefore, paired t -tests were performed for each acoustic feature comparing pairs of affects to determine which features were significantly different between each pair.

4.1. Overall pitch

The mean and range of F_0 are two popular features that have been reported on in many studies. Banse and Scherer [1] summarize previous findings as follows. Affects involving high arousal levels such as anger, fear, and happiness are characterized by an increase in F_0 mean and range whereas sadness is characterized by a decrease in F_0 mean and range. Cahn [3] reported a similar trend for F_0 range, but for F_0 mean her findings were much different in that fear showed the highest contribution followed by sad, then happy and angry. Figure 3 shows the mean differences between the affect pairs and the 95% confidence intervals for the F_0 mean for our speaker. The t -values and p -values were obtained by performing the paired t -tests. The F_0 mean values for this recording set were 279 Hz

for happy, 261 Hz for angry, 250 Hz for fearful, and 177 Hz for sad. The sad affect is significantly lower in pitch than the other three emotions, in line with previous studies. Happy is slightly higher than fearful. The differences between angry and happy and between angry and fearful are not significant. The F_0 range shows the same picture as the F_0 mean in terms of the differences between the affect pairs. The average F_0 range is 581 Hz for happy, 544 Hz for angry, 431 Hz for fearful, and 309 Hz for sad. Note that these are recordings from a child, which explains the high range in F_0 . All F_0 range differences between affect pairs are significant, except the difference between angry and happy.

The F_0 mean and range are not very informative features for describing the pitch contours. Using parameters derived from the phrase curves and accent curves as obtained from our decomposition algorithm, allows for a more detailed description of the differences between affects.

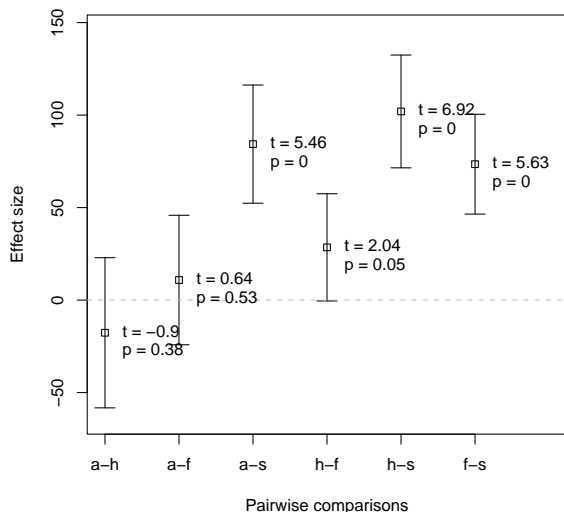


Figure 3: F_0 mean differences between affects.

4.2. Phrase curves

Due to the shortness of the sentences, there were no minor phrase boundaries and as such there was only one phrase curve per sentence. Anger and fear have been found to have more declination than happy and sad [1], although in a different study anger and sad were found to have a level contour slope and happy and fear had a rising contour slope [3]. The problem with these analyses is that they derive the declination slope from the raw pitch contour, the slope of which is polluted by the pitch accent prominences. The main advantage of our decomposition algorithm is that it allows for a separation of the declination in the phrase curve from the accent curves. Figure 4 shows differences in the average phrase curve range, which is defined as the difference between the maximum and the minimum value of the phrase curve. The results show that the differences in phrase curve range between angry and happy and between fearful and sad are not significant. However, both angry and happy have a significantly larger range than fearful and sad. The average phrase curve range is 188 Hz for happy, 200 Hz for angry, 120 Hz for fearful and 90 Hz for sad.

We also computed the average slope of the phrase curve (or declination). The results show the same trends as for the

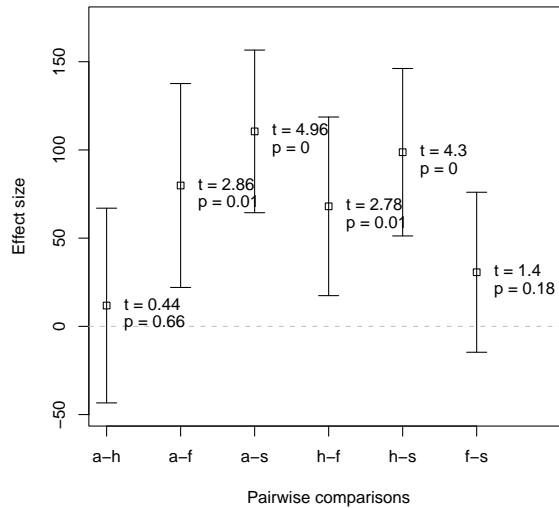


Figure 4: Average phrase curve range differences between affects.

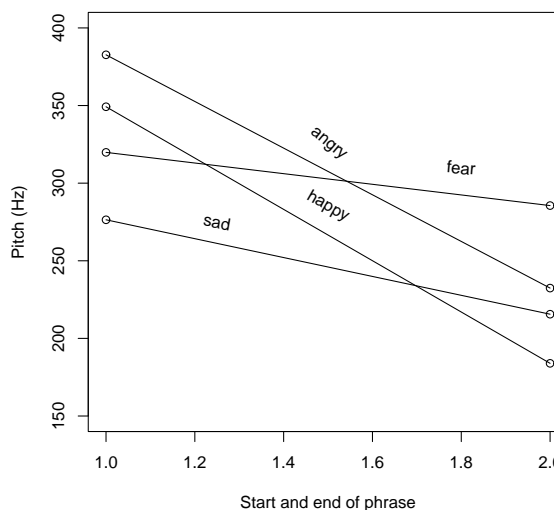


Figure 5: Average phrase curve start and end values for each affect.

phrase curve range differences in that the differences between angry and happy and between fearful and sad are not significant. However, both angry and happy have significantly less declination than fearful and sad. The average slope of the phrase curve is -1.74 units for angry, -1.45 for happy, -0.29 for fearful and -0.71 for sad. The phrase curves for the fearful condition are almost flat.

Figure 5 displays the average start and end points of the phrase curve for each affect. The difference in slope is clearly visible between on the one hand the angry and happy and on the other hand the fearful and sad affects. The slope difference is mainly related to the end point of the phrase curve. The phrase curve on average starts higher for the angry affect than for happy, followed by fearful and sad. But the phrase curve ends highest for fear, followed by angry, sad, and happy. These findings will be very helpful for applying appropriate phrase curves to the phoneme sequences in our recombinant synthesis system.

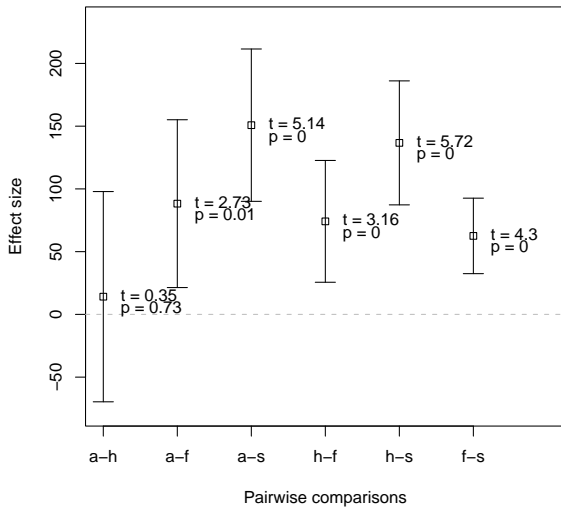


Figure 6: Average accent curve height differences between affects.

4.3. Accent curves

The start of the accent curve always coincides with the start of the foot, which is always a stressed/accented syllable. The end of the foot is located at the end of an unstressed syllable either right before the start of the following foot, or a phrase boundary. However, previous research has shown that the end of the accent curve does not need to coincide with the end of the foot, leading to overlapping accent curves [8]. We were able to provide a satisfactory fit to the pitch contours using accent curve templates for the basic up-down shape, negative accents and accents with continuation rises. We found some negative accents in our corpus, but the occurrence of negative accents was not significantly different between affects. Because the sentences were so short, there were no minor phrase boundaries and thus no continuation rises at those locations. But the speaker would sometimes end sentences in a continuation rise. Our hypothesis was that this occurred mostly in the fearful and sad affects, but no significant effect was found. For the measurement of accent curve amplitudes, the negative accents were excluded from the analysis.

Figure 6 displays the average differences in accent curve amplitudes between the affect pairs. The accent curve amplitude is measured at the peak location. It can be observed that the difference in accent curve amplitudes is not significant for the angry-happy comparison, but it is significant for all other comparisons. Both angry and happy have higher accent amplitudes than fearful and sad. Fearful has higher accent curve amplitudes than sad. The average values for the four affects are: 172 Hz for angry, 173 Hz for happy, 77 Hz for fearful and only 27 Hz for sad.

For sentences that had more than one accent, we also studied the average accent curve amplitude for the first accent versus that of the last accent. The averages are based on 60 out of 96 sentences. The first peak was on average 133 Hz for angry, 176 Hz for happy, 76 Hz for fearful and 29 Hz for sad. For the last peak the average values were 157 Hz for angry, 181 Hz for happy, 93 Hz for fearful and 18 Hz for sad. This shows that for all conditions except sad, the final accent had a higher amplitude than the first one.

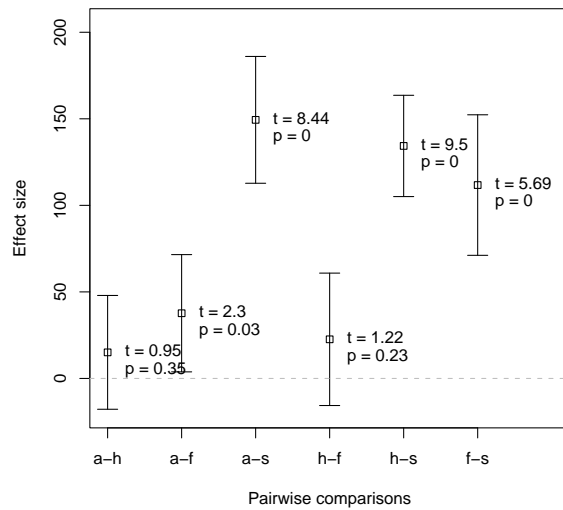


Figure 7: Average overall energy differences between affects.

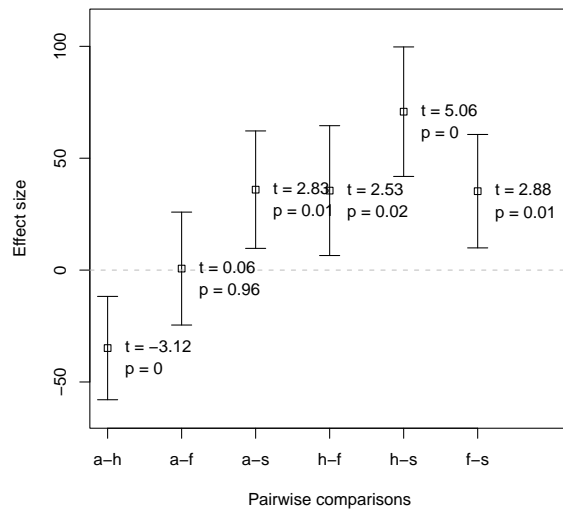


Figure 8: Average spectral tilt differences between affects.

4.4. Energy

Figure 7 shows the overall energy differences between affect pairs. The overall energy was computed as the sum of the four broad spectral band averages. As can be seen, the overall energy for sad is much lower than for the other three affects. Fearful is significantly lower than angry but its lower overall energy with respect to happy is not significant. Angry is louder than happy but again this difference is not significant. The average overall energy for angry is an order of magnitude of 409 for angry, 394 for happy, 372 for fearful and 260 for sad.

Although spectral tilt was not found to be a significant factor using the analysis of variance, we do include it here, as the paired *t*-test showed that there was an important difference in spectral tilt between angry and happy. This makes the spectral tilt one of the few parameters to distinguish angry from happy in our corpus. Figure 8 displays the average spectral tilt differences between affect pairs. The most important finding is that

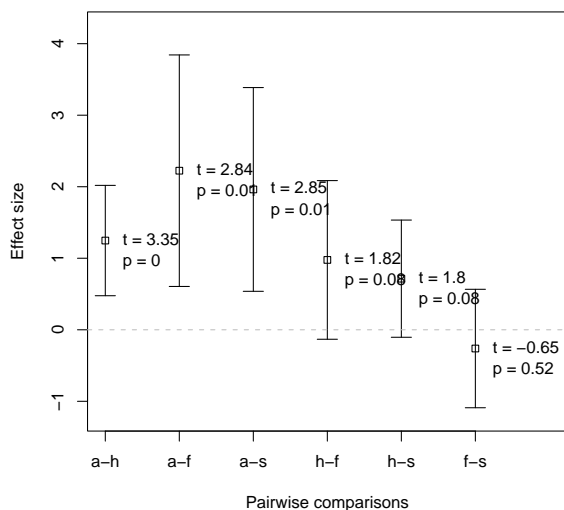


Figure 9: Average speaking rate differences between affects.

the spectral tilt in anger is significantly lower than in happy. The average values for spectral tilt were -87 units for angry, -53 for happy, -88 for fearful and -123 for sad. Thus, sad has the lowest amount of high-frequency energy whereas the other three emotions, all three of which are associated to higher arousal levels according to Banse and Scherer, have higher amounts of high-frequency energy, which is reported to be due to an increased vocal effort by the speaker [1].

4.5. Speaking rate

Phoneme durations and pause lengths are often included in an analysis of different affects. Because the sentences in our corpus are relatively short, there are no intermediate pauses that can be analyzed. We computed the average speaking rate by dividing the total phoneme duration (excluding pauses) by the number of phonemes. The average speaking rate was 140 ms/phoneme for angry, 127 ms/phoneme for happy, 117 ms/phoneme for fearful and 116 ms/phoneme for sad. This is surprising as we expected the angry affect to be faster than the other affects, but for this speaker that turned out not to be the case. We also considered other duration measures such as vowel durations and voiced portion durations, but the effects were similar to the speaking rate findings, so we don't go into detail here.

5. Conclusion

The sad affect presents the most distinct acoustic and prosodic features from the other three affects. The sentences have a lower overall energy and higher spectral tilt. The phrase curves are lower and the accent curve amplitudes are much lower than in other affects. The other three affects (angry, happy and fearful) are all high-arousal emotions and can be more easily confused with each other. However, our analysis has shown that we can distinguish the three affects for our speaker. Fearful is distinguishable from angry and happy by showing a lower F_0 range, a flatter phrase curve and lower accent curve amplitudes. Angry is distinguishable from happy by displaying a higher spectral tilt and a slower speaking rate.

The results provide a promising start to synthesizing expressive speech using our recombinant synthesis approach. The decomposition algorithm was shown to do a good job decom-

posing the pitch contours into phrase and accent curves, despite the fact that we were dealing with highly expressive children's speech. This demonstrates the fact that a prosodic corpus using neutrally read sentences can be used to select phrase and accent curves, which can then be warped using different warping functions for each affect to exhibit varying phrase curve slopes and ranges and varying accent curve amplitudes. The phonemic units selected from the acoustic corpus can be warped in the sinusoidal framework to display varying overall energy and spectral tilt profiles using the four-band representation.

6. References

- [1] R. Banse and K. Scherer, "Acoustic Profiles in Vocal Emotion Expression", In *Journal of Personality and Social Psychology*, 70(3), pp. 614-636, 1996.
- [2] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer", [online <http://www.fon.hum.uva.nl/praat/>]
- [3] J. Cahn, "Generating Expressions in Synthesized Speech", Master's Thesis, MIT, 1989.
- [4] J. P. Hosom, "Automatic Time Alignment of Phonemes using Acoustic-Phonetic Information", PhD Thesis, Oregon Graduate Institute, Beaverton, OR, 2000.
- [5] E. Klabbbers and J. van Santen, "Control and prediction of the impact of pitch modification on synthetic speech quality", In *Proceedings of EUROSPEECH'03*, Geneva, Switzerland, pp. 317-320, 2003.
- [6] E. Klabbbers and J. van Santen, "Expressive speech synthesis using multilevel unit selection (A)", In *J. Acoust. Soc. Am.* 120(5), pp. 3006, 2006.
- [7] Q. Miao, X Niu, E. Klabbbers, and J.P.H. van Santen, "Effects of Prosodic Factors on Spectral Balance: Analysis and Synthesis", *Speech Prosody 2006*, Dresden, Germany.
- [8] T. Mishra, J.P.H. van Santen, and E. Klabbbers, "Decomposition of Pitch Curves in the General Superpositional Intonation Model", *Speech Prosody 2006*, Dresden, Germany.
- [9] S. Mozziconacci, "Speech Variability and Emotion: Production and Perception", PhD Thesis, Technical University Eindhoven, 1998.
- [10] J. van Santen and B. Möbius, "A quantitative model of F_0 generation and alignment", In A. Botinis (ed.), *Intonation: Analysis, Modeling, and Technology*, pp. 269-288, Kluwer Academic Publishers, Netherlands, 1999.
- [11] J. van Santen and X. Niu, "Prediction and Synthesis of Prosodic Effects on Spectral Balance of Vowels", 4th *IEEE Workshop on Speech Synthesis*, Santa Monica, CA, 2002.
- [12] J. van Santen, T. Mishra, and E. Klabbbers, "Estimating phrase curves in the general superpositional intonation model", In *Proceedings of the ISCA Speech Synthesis Workshop*, Pittsburgh, PA, 2004.
- [13] J. van Santen, A. Kain, E. Klabbbers, and T. Mishra "Synthesis of prosody using multi-level sequence units", *Speech Communication*, 46(3-4), pp. 365-375, 2005.
- [14] K. Scherer, "Vocal communication of emotion: A review of research paradigms", *Speech Communication*, 40, pp. 227-256, 2003.