Pitch modulation in emotional speech by non-emotive factors: case studies

Irina Gorodnitsky [*]

Department of Cognitive Science, University of California, San Diego, 9500 Gilman Dr.,
La Jolla, CA 92093-1505, USA
Phone: 858-822-3221; Fax: 858-534-1128

Abstract

Automatic affect recognition (AAR) from speech is often investigated ignoring extraneous, non-emotive prosodic information. In reality, spontaneous speech contains a multitude of non-emotive cues that modulate pitch/F0. This paper presents an investigation of how stress placement influences utterance duration and aggregated F0 measures - mean, minimum, and range - that are used in AAR. The second part of the analysis examines whether we can identify stress placement in emotive speech from F0 contours sufficiently reliably that it can be accounted for in AAR. The investigation uses sample utterances from an emotional speech corpus, which includes as variables three emotion categories (neutral, angry, and happy) crossed with two placements of stress. The lexical content of the sentences also varied, either matching or mismatching the emotional expression, and this condition was also included in the analysis. Results show that stress placement distorts the relationships between the emotional categories and the measures that include aggregated F0 and utterance duration. Likewise, the expressed emotions influenced contour shapes so that they did not correlate in a reliable way with accentuation. In contrast, there was little evidence that lexical content influenced any of the results.

Key words: Speech, Pitch, Prosody, Emotion, Automatic Affect Recognition,
Human-Machine Interfaces, Nonlinear oscillator

1 Introduction

Emotions are a major aspect of human experience. Human-computer interaction (HCI) practitioners have been advocating for inclusion of emotions in the inter-

[*] Corresponding author.
Email address: igorodni@ucsd.edu (Irina Gorodnitsky).
URL: www.cogsci.ucsd.edu/∼igorodni (Irina Gorodnitsky).

faces, both in the design of man-made systems Norman (2004) and in assessing a system's effectiveness Picard (1997). For these interfaces, vocally expressed emotions are most frequently used Picard (1997). Despite the recognized importance of affect in communication, many key questions remain regarding mapping of expressed emotions in human-machine interfaces (HMI). A range of perspectives exists on what we can or should measure in affective speech. Perspectives vary on such key issues as whether emotions should be classified in a discreet space Ekman (1999) or in a continuous multi-dimensional domain, e.g. Lang (1995). There is also a question of how prevalent vocally expressed emotions are in everyday speech Cowie and Cornelius (2003), which brings up the pragmatic issue of the usefulness of emotions over 'moods' in practical interfaces. The research in Cowie and Cornelius (2003) and references therein, found not only a surprising lack of expression of basic emotions in everyday speech, but also that *fully blown emotions*, the types many automatic affect recognition (AAR) systems attempt to identify, are extremely rare in speech. This extended even to situations where people were discussing emotional life events (Cowie et al., 1999).

In additions to these debates on what we can measure, there are also serious methodological issues in the computational domain which have to do with the identification and evaluation tools for spoken affect. Even within the limited scope of recognizing only the primary emotions (i.e. anger, disgust, fear, happiness, sadness, and surprise (Cornelius, 1992)), the task remains formidable and as a result, modern AAR systems are quite complex. The systems use a multitude of features, the majority based on F0/pitch, that include F0 mean, standard deviation, range, mean duration, variability, slope, and jitter (number of changes in sign of the F0 derivative) (Iida et al., 1998; Heuft et al., 1996; Noad et al., 1997; Dellaert et al., 1996; Amir and Ron, 1998). These features are augmented by non-F0 features, the most common of which are the mean energy of a speech signal, value of high frequency energy, energy per phone, articulation rate, speech rate (frequency of occurrence of unvoiced periods), speed of speaking (duration of inter-sentence silences), intensity, intensity variance and tremor (measure of tremor in the intensity over the intensity curve). The large number of degrees of freedom defines a complex classification space that is managed using non-trivial scoring functions. There are several successful approaches that use a hierarchal system, e.g. the features are divided into long-term (high-level) and short-term (low- level). Depending on the classification task and the degree of speaker dependent training, AAR rates ranging from 55% to 90% have been reported when using such systems, i.e. (Li and Zhao, 1998; Noam, 2001; Noquerias et al., 2001).

The quoted rates in the high end of the range, however, were achieved in non- adverse recoding conditions and for very restrictive (i.e. binary) classification tasks. In more realistic scenarios the rates are considerably lower. Noise has a large negative impact on AAR, since even moderate noise levels dramatically worsen the accuracy of F0 tracking. High-dimensionality of the classification space is another factor that negatively impacts AAR. Yet, a third factor exists, which has not gained much attention. That is the influence of prosodic features other than emotions on pitch and, by extension, on AAR success. Pitch is the essential element used in AAR, but it also bears a direct relationship to linguistic production and is a key feature of *all* aspects of prosody, not just affect. Pitch has been established as the vocal cue linked to emphasis, stress, irony, truthfulness, intent, attitude, temperament, laughter, and other paralinguistic properties. It is reasonable to suspect that modulation of pitch by these prosodic factors could diminish the efficacy of AAR. It should be noted that even though there is a tendency to use pitch and F0 interchangeably in literature, they are not identical but are closely related. Pitch refers to the perceived fundamental frequency of sound while F0 describes the fundamental frequency of the produced speech and is the quantity that is estimated by most algorithm.

To the author's knowledge, a few high quality emotional speech corpuses exist, but none in public domain, that incorporate non-emotive cues that are independent from the emotional categories. For this paper, sample utterances were used from one such emotional speech corpus described in (Alter et al., 2003). The corpus includes three emotional categories - neutral, angry, and happy - crossed with two placements of stress, resulting in 6 conditions. Since stress placement is known to influence intonation, the treatment of stress placement independent from the emotions is an essential feature of this corpus. Another potentially relevant condition included in the corpus is a matched/mismatched lexicon-affect condition. Each sentence in the corpus carries positive, negative, or neutral connotation. The connotation of each sentence was classified by a group of experimental subjects and the sentences were recorded spoken with three emotive styles.

The central part of the analysis examines how the presence of an extra prosodic variable, namely, the stress placement, affects the aggregated F0 measures utilized in AAR: mean, range, and the global minimum value, and also temporal, non-F0 measure that is utterance duration. In addition, the second part of the analysis examines whether stress placement and the shape of the F0 contour are reliably correlated in emotional speech. The goal of this part of the analysis is to find whether stress placement in emotional speech could be automatically recognized and used as a constraint in AAR. Lastly, effects of the mismatched versus matched lexicon-affect on F0 contours are also examined.

Opinions are abundant on how intonation (F0 contours) should be modeled. The formalized models of intonation come from linguistics. However, in dealing with emotional speech there are many methodological and theoretical limitations imposed by the linguistic models, as discussed in great detail in Bnziger and R. (2005). Discussing these issues is outside the scope of this paper, but it is important to note that linguistic models have not been researched in the context of emotional speech, with the sole exception of Mozziconacci and Hermes (1999). Thus, it is unknown how emotional categories influence features of such models, which raises uncertainty regarding the applicability of these models to the analysis of F0 contours here. It is clear that the categorical changes in F0 contours used in such models do not apply well to the continuous variations in F0 contours caused by voiced emotions.

An alternative to the linguistic models was suggested in Bnziger and R. (2005) that accounted for continuous variations in F0 contours. The system in Bnziger and R. (2005) used points that marked fixed features across F0 contours. The operationally defined ’accents’ in the utterances, which are defined by a minimum, maximum, and minimum of an F0 excursion, were used for the analysis. This system provided a practical way for quantifying the relationship between intonation patterns and emotions and it accommodated well the utterances used in Bnziger and R. (2005), which contained no syntactic or semantic information. This system did not however prove to be practical for the analysis of stress placement reported here. Part of the reason is that we are looking at the combined effect of two prosodic elements - the stress placement and emotion. The combination produced large variations in the size, shape, and pattern of the ’accents’. Also, unlike the meaningless sequences of syllables used in Bnziger and R. (2005), the utterances here are meaningful sentences. Measures used in any analysis must ultimately tie to the goals of the study. After examining the available methods for modeling F0 contours, the author concluded that relating variations in ’accent’ patterns to the stress placement, as presented in section 3, was the most relevant to the goals of the current analysis.

The aim of the presented work is to spotlight the potential impact non-emotive features may have in AAR. It should be noted that the author did not collect the corpus used here, and only 42 utterances from that corpus were available to the author, representing the total of seven sentences and 18 categories (three emotions crossed with two accentuations and with 3 lexical categories). Given the large number of dependent variables and the small sample size, aiming for statistically significant analysis was not meaningful in this study. Even in analyses of much larger corpuses, as for example in Bnziger and R. (2005), the combination of multiple conditions leads to an explosion in the number of categories such that the small sample size within categories affects the statistical significance of the results. Rather, the results presented here should be viewed as suggestive of the issues and directions for future statistically rigorous studies. It is also important to note that practical interfaces must deal with individual speech cases and not average trends. In that regard, case studies are invaluable because they provide ’cases in point’ demonstrations that bring awareness to issues that must ultimately be dealt with.

2 Methods

2.1 Emotional Speech Corpus

The experiments in this paper use seven sentences selected from the German language emotional speech corpus described in (Alter et al., 2003). This corpus contains sentences with identical syntactic form (subject-auxiliary-NP-verb), where NP stands for ’the nominal phrase’. The emotional meaning of each sentence in the database was classified by experimental participants (n=20) who rated the written sentences using three categories: neutral, positive, or negative. The sentences were then recorded while spoken in Standard German by a trained female actress. Each sentence was spoken and appeared in the database exactly 6 times, recorded using six different expressive styles. These different styles were created using three forms of emotional expression: a neutral speaking style and two types of primary emotions (happiness and cold anger) crossed with two placements of stress: on the NP and on the final verb. Thus the recorded utterances either matched or were at odds with the sentence’s lexical content.

Seven sentences, which were numbered 17, 43, 83, 85, 112, 123 in the corpus, were made available to the author. These sentences were chosen at random and all six recordings of each sentence (three emotional categories and two stress placements) were provided, for the total of 42 records. Of the seven sentences, three were rated as having positive lexical content, three sentences rated as having negative lexical content and one sentence rated as having neutral content. The sentences are listed below. The first number indicates the sentence order referenced in the analysis, the second number corresponds to the numbering used the in the original corpus and the sign ’+’, ’-’, or ’0’ indicates the positive, negative, or neutral emotional rating of the sentence as assigned by the experimental participants.

1 17 − Sie hat ihn mit der Waffe bedroht.

(She threatened him with the weapon.)

2 40 − Er hat sie von der Klippe gestoben.

(He pushed her off the cliff.)

3 83 − Er hat ihn ins Gesicht geschlagen.

(He slapped him in the face.)

4 85 + Sie hat es ans Licht gebracht.

(She brought some facts to light.)

5 103 + Er hat um ihre Hand angehalten.

(He asked for her hand in marriage.)

6 112 + Sie hat den Rekord gebrochen.

(She broke the record.)

7 123 0 Er hat den Brief geschrieben.

(He wrote the letter.)

It should be noted that one of the original objectives in creating this corpus was to examine the connection between affect-dependent acoustic features and the neural responses of listeners, which were monitored using event-related brain potentials (ERPs). The analysis in the present study uses only the speech corpus.

2.2 Analysis Methods

The recordings in (Alter et al., 2003) were done at 44100Hz. F0 for the voiced segments was computed using a nonlinear oscillator (ID) method (Gorodnitsky, 2007). The performance of the ID method is the same as that of the conventional pitch estimation algorithms in noise free conditions, but in medium to high noise ID appears to offer greater robustness (Gorodnitsky, 2007). F0 computed using the ID method were further spot checked using ESPS algorithm on 5 randomly selected utterances. The ESPS implementation was obtained from the Snack library audio analysis module for Linux (Sjlander, 2005). ESPS computes pitch using the normalized cross correlation function and dynamic programming. Both methods were run with a 100-400Hz limit on F0 and 3 msec and 10 msec frame spacing (framelength in SNACK) and window size respectively. Identical F0 estimates were obtained with both methods for the noise-free utterances.

3 Results

In the first part of this section we examine how the placement of stress in speech may interfere with the analysis of F0 related to expressed emotions. The following analysis highlights instances of this interference when the stress is placed on the final verb rather than the nominal noun. As discussed in section 1, confidence intervals on the observed correlations are too large to support rigorous statistical analysis of their significance, because of the small number of utterances in each dependent condition. The point then of this study is to identify apparent correlations between measures of F0 and the prosodic conditions that suggest areas for further investigation.

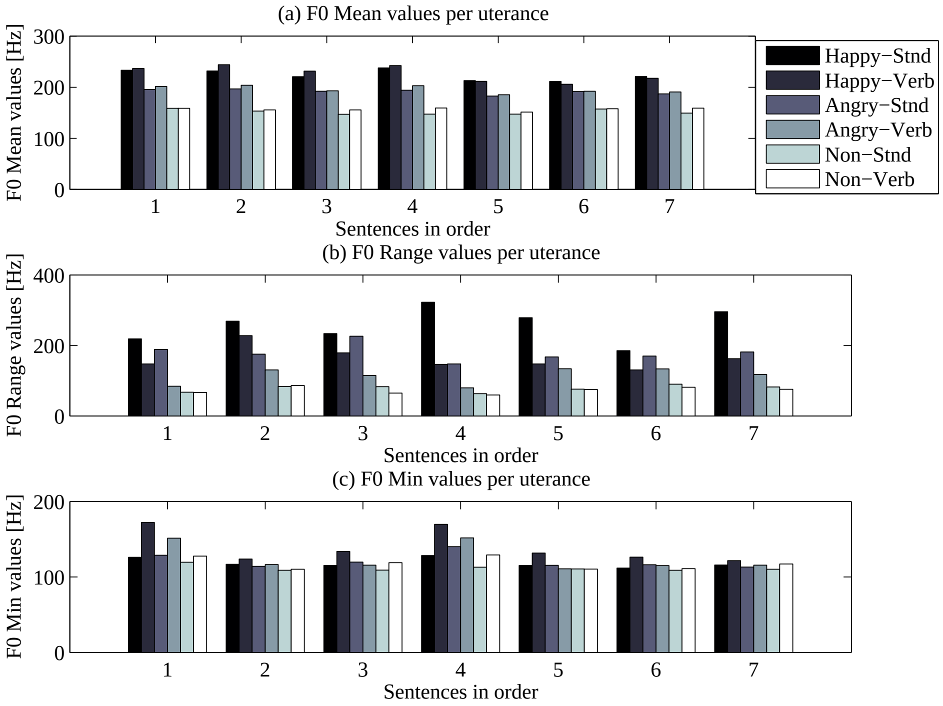

Fig. 1. Aggregated measures of F0 - F0 mean, F0 range, and F0 min - for the six prosodic conditions: ’happy’ expression, standard stress placement on the noun (HAS); ’happy & stressed verb’ (HAV); ’angry & standard stress on the noun’ (ANS); ’angry & stressed verb’ (ANV); ’unemotional & standard stress on the noun’ (NOS); and ’unemotional & stressed verb’ (NOV). The legend for all plots is shown in (a).

Aggregated F0 measures: Fig. 1 shows aggregate F0 descriptors - F0 mean, F0 minimum, and F0 range - for the six prosodic conditions which are produced from the combinations of the three expressive styles and two types of accentuation: ’happy & standard stress on the noun’ (HAS), ’happy & stressed verb’ (HAV), ’angry & standard stress on the noun’ (ANS), ’angry & stressed verb’ (ANV), ’unemotional & standard stress on the noun’ (NOS), and ’unemotional & stressed verb’ (NOV).

F0 mean (Fig. 1a): There is an overall correlation between F0 mean and the expressed emotion, where the ’happy’ expressions are associated with the highest F0 mean while non-emotive cases with the lowest, with the ’angry’ cases in the middle. However, the location of the stress modulates F0 mean. Compared to the stress on the noun (NS), F0 mean is raised in each emotional category when the stress is on the verb (VS). Sentence #6 ’happy’ condition, where F0 mean decreases, is the only exception). The increase is not enough to change the correlation pattern between F0 mean and the three expressed emotions. However, it introduces a consistent bias which adds to the uncertainty in AAR classification brought about by other factors, such as noise.

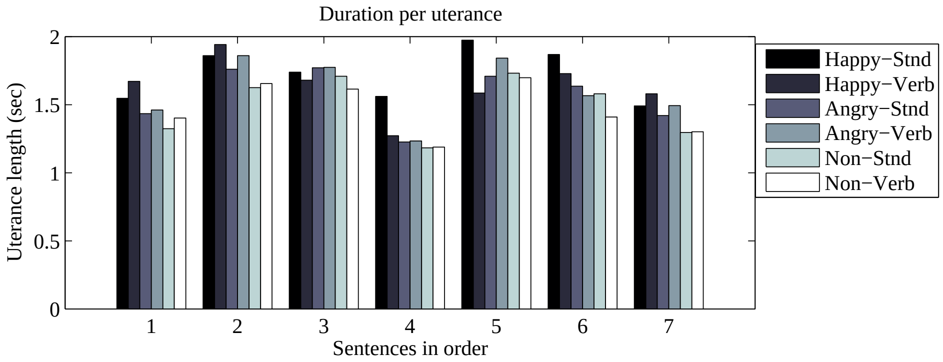

Fig. 2. Duration of the utterances for the six prosodic conditions: ’happy’ expression, standard stress placement on the noun (HAS), ’happy & stressed verb’ (HAV), ’angry & standard stress on the noun’ (ANS), ’angry & stressed verb’ (ANV), ’unemotional & standard stress on the noun’ (NOS), and ’unemotional & stressed verb’ (NOV).

F0 range (Fig. 1b): F0 range shows a pattern of correlation with emotional categories closely resembling that for F0 mean, that is the ’happy’ expressions are associated with the highest F0 range while non-emotive cases with the lowest, with the ’angry’ cases in the middle. However, there are numerous exceptions where this correlation pattern does not hold. In most cases, the exceptions are associated with stress is on the verb (VS), which reduces the F0 range relative to F0 range for NS. The break in the correlation pattern can be observed, for example, in HAV versus ANS condition in six out of the seven sentences (# 1,3,4,5,6,7).

F0 min (Fig. 1c): Although F0 min is used in AAR, it does not have a strong correlation pattern with the emotive expressions in our examples. However, it is also affected to a great degree the by stress placement. For example, note that for NS, F0 min values are consistently lower for non- emotive versus emotive expressions. VS consistently raises F0 min values for non-emotive cases such that the F0 min values are equal or exceed the F0 min values for the emotive cases with stress on the noun, thus erasing the potential distinction between the emotive and non-emotive expression in the F0 min measure. This occurred in sentences # 1, 3, 4, 6 in the ’happy’ case and in sentence #7, in both ’happy’ and ’angry’ cases. This inconsistency in F0 min values in VS versus NS condition would bias AAR, adding to the biases brought about by the inconsistencies in F0 mean and range values discussed above.

Temporal measures: Utterance duration as well as the duration of segments within utterances provide another important set of AAR measures related to articulation rate and speech rate. Fig. 2 illustrates the duration of the utterances as a function of the same six prosodic conditions used in Fig. 1. The general pattern seen in Fig. 2 shows that the ’happy’ expression has the longest duration utterances while the non-emotive case has the shortest, with the ’angry’ expression in between. Importantly, as in the case of aggregated F0 measures, VS alters the duration of the utterances, but in an inconsistent way across the three emotional categories, introducing a similar bias to AAR. Whether the duration shortens or lengthens with the change of stress to the verb depends on the expressed emotion. VS lengthens the duration overall in the three negatively rated sentences and the neutrally rated sentence. On the other hand, VS mostly shortens the duration in the three positively rated sentences. In the cases where the duration lengthened due to VS, in 50% of these cases, sentences # 2 and 7, the length differences between the emotional categories: 'happy' (HAS) and the 'angy' (AVN) expression were completely erased. As another example, in sentence $\#1$, the lengthening due to VS in the non-emotive condition made the utterance duration almost equal to that of the standard stress 'angry' (ANS) condition.

More detailed analysis of the effect of stress placement on individual voiced segments within the utterances was also performed. It showed a very slight tendency toward the voiced segments being shorter when the stress was placed on the verb. Since each sentence contains several voiced segments, it provided us a larger sample size to work with than the utterances on the whole. Hence, ANOVA was computed to analyze the effect of stress placement on the duration of voiced segments, but no statistical significance was found. It should be noted that the length and continuity of the voiced segments relates to the sustained vocal effort which a trained actress is expected to express maximally in each utterance, as she is expected to annunciate every sound. A related observation was that the speaker tended to place greater vocal effort prior to and during the accented point. After the accent point F0 contours sometimes became less well defined as can be observed from Fig. 3 and 4. These factors would be expected to affect the results of the analysis here.

In summary, the shift of stress to the final verb was found to be associated with a number of changes in the aggregated F0 and temporal measures commonly utilized in AAR. Since the current AAR systems do not account for stress placement, it is reasonable to expect that such changes would negatively impact AAR success. Hence, the main goal of this study - spotlighting the possible bias that stress placement can bring to AAR - has been fulfilled.

One can also ask a follow-up question of how feasible it may be to anticipate and ameliorate effects of stress placement in AAR. As was explained in section 1, we study this question by examining the 'accents' pattern as a function of the stress placement in the utterances. As part of this analysis we first look at whether lexical content may also influence F0 contour shape. Fig. 3 compares pitch values for the three lexically different sentences (# 2 (40), 5 (103), and 7 (123)) across the two emotive cases, happiness and cold anger, and two stress placements. Note: the sentence for each lexical subcategory used in this section of the paper was chosen at random. The examples discussed throughout this section are representative of the results observed for all the sentences, unless explicitly stated otherwise.

The following analysis examines locations of 'accents', which are characterized by the mid-range rise and fall of F0 over 100 to 400 msecs. We can see from Fig. 3 that even though we are comparing contours for completely different sentences

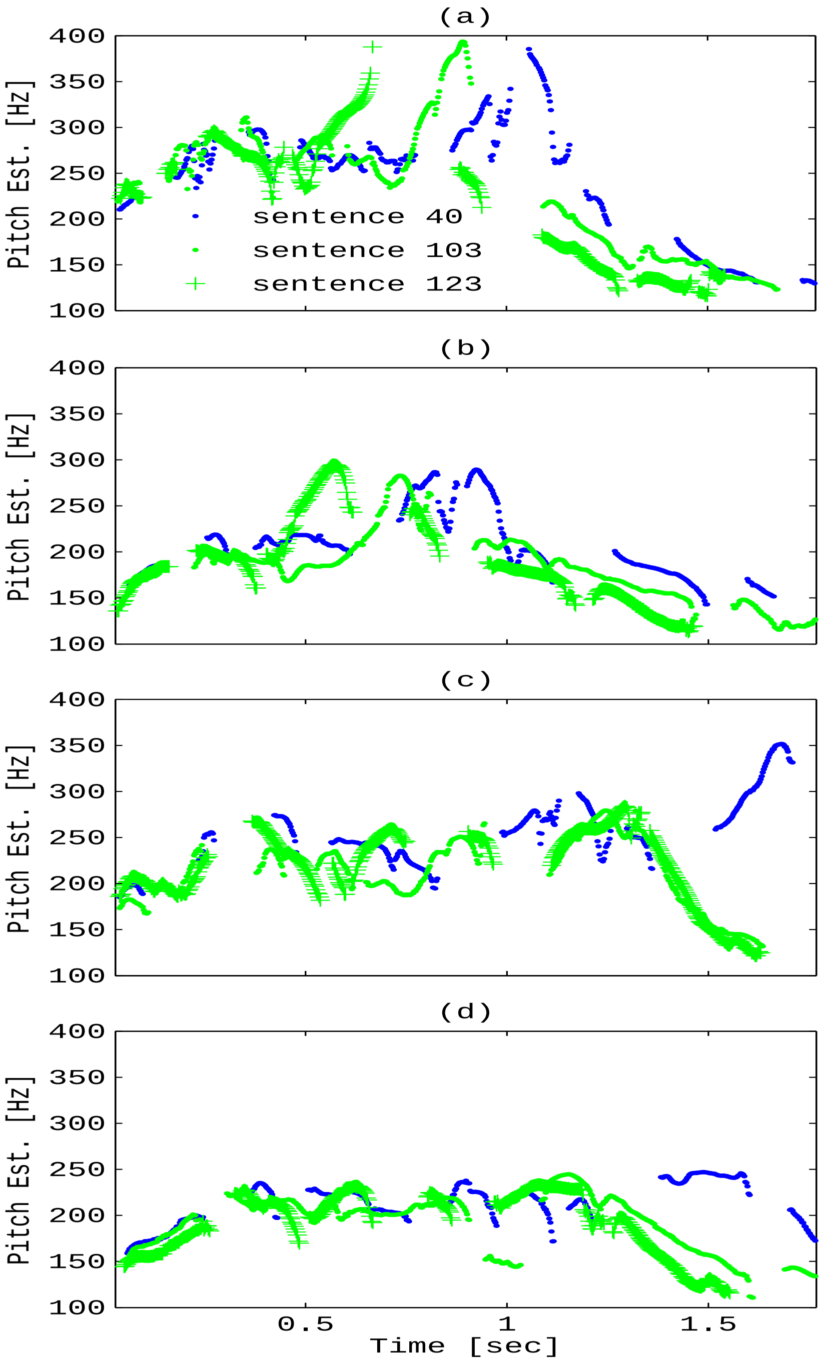

Fig. 3. F0 contours for three lexically different sentences separated by two expressed emotions and stress placement: (a) expression of ’happiness’, stress on the noun. (b) expression of anger, stress on the noun. (c) expression of ’happiness’, stress on the final verb. (d) expression of anger, stress on the final verb. The legend for all plots is shown in (a) and refers to the sentence numbers used in the original corpus.

there is good general agreement between the gross long-range F0 features for the three sentences in the NS case. The ’accents’ appear at different times in the three sentences shown in Fig. 3a and b, but in each case they coincide with the stressed noun. Hence, adjusted for differences in the length of the utterances, the F0 contours in Fig. 3a and b differ only in minor respects from each other and the stress placement is correlated with the F0 maximum. Sentence $\#2(40)$ presents a minor exception from this rule because its 'accent' is split into two peaks. This appears most prominently in the ANS case (Fig. 3b). In the VS case, the agreement between the gross long-range F0 features for the three sentences is not as clear as in the case of NS. F0 contours in sentence $\#2(40)$ in Figs. 3c and d clearly deviate from the other two at the point of accentuation, rising while the other two fall. It is the only sentence in which the rise correlates with VS. For the other two sentences, VS either has an weakly defined 'accent' associated with it's location in the utterance (Fig 3c) or does not have an 'accent' at all at it's location (Fig. 3d). By weakly defined 'accent' we mean an accent that has its maximum equal to or smaller than the global maximum of the F0 contour. Sentence $\#2(40)$ shows that lexicon could possibly influence F0 contour shape independently from the prosodic conditions, although the influence in our example is seen only in the case of nonstandard accentuation. However, extraneous factors that we do not account for here could also have been responsible for the unusual F0 contour shape seen in the single example of sentence $\#2(40)$.

Next, we examine the relationship between the stress placement and the pattern of 'accents' for the three emotive conditions. The examples in Fig. 3 discussed above already have shown that VS does not necessarily have an associated strong 'accent' that can be used for recognition of stress. In Fig. 4, we show F0 contours for the three emotive conditions, differentiated by the stress placement and lexicon. The contours begin at about the same level, but separate quickly and settle at a distinct, well-defined F0 level for each emotive condition, as expected. Note that in the NS case, for the neutral emotional category (Fig 4 a, c, and e, bottom curves), F0 varies very little throughout and there are no "accents" at NS. The lexically neutral sentence is the only one in this example where NS placement coincides with a prominent 'accent' in the utterances spoken with the two emotions. In the positively rated sentence $\#4(85)$, both F0 reach their maxima at the end of the sentence and NS does not coincide with an 'accent'. In the negatively rated sentence $\#1(17)$, the utterances spoken with emotion do have 'accents' coincident with NS, but the F0 contour in the HAS case has multiple accents which make it harder, if not impossible to recognize NS based on the F0 maxima. In the case of VS, we again observe a variety of contour patterns. With one exception (HAV, sentence $\#4(85)$, no utterance has the largest 'accent' associated with VS. We observe multiple, often equal sized accents throughout the utterances for every emotional category. In the HAV case, sentence $\#1(17)$, there are two equally sized maxima, one at the noun and the other at the final verb. The conclusion here is that stress placement does not have a set of associated long-range F0 features in the given utterances.

The effect of stress on aggregated F0 measures - mean, range, and minimum - is also apparent in Fig. 4. F0 contours maintain distinct ranges only prior to the stress point. After that point, in many cases, F0 curves for the different emotive categories

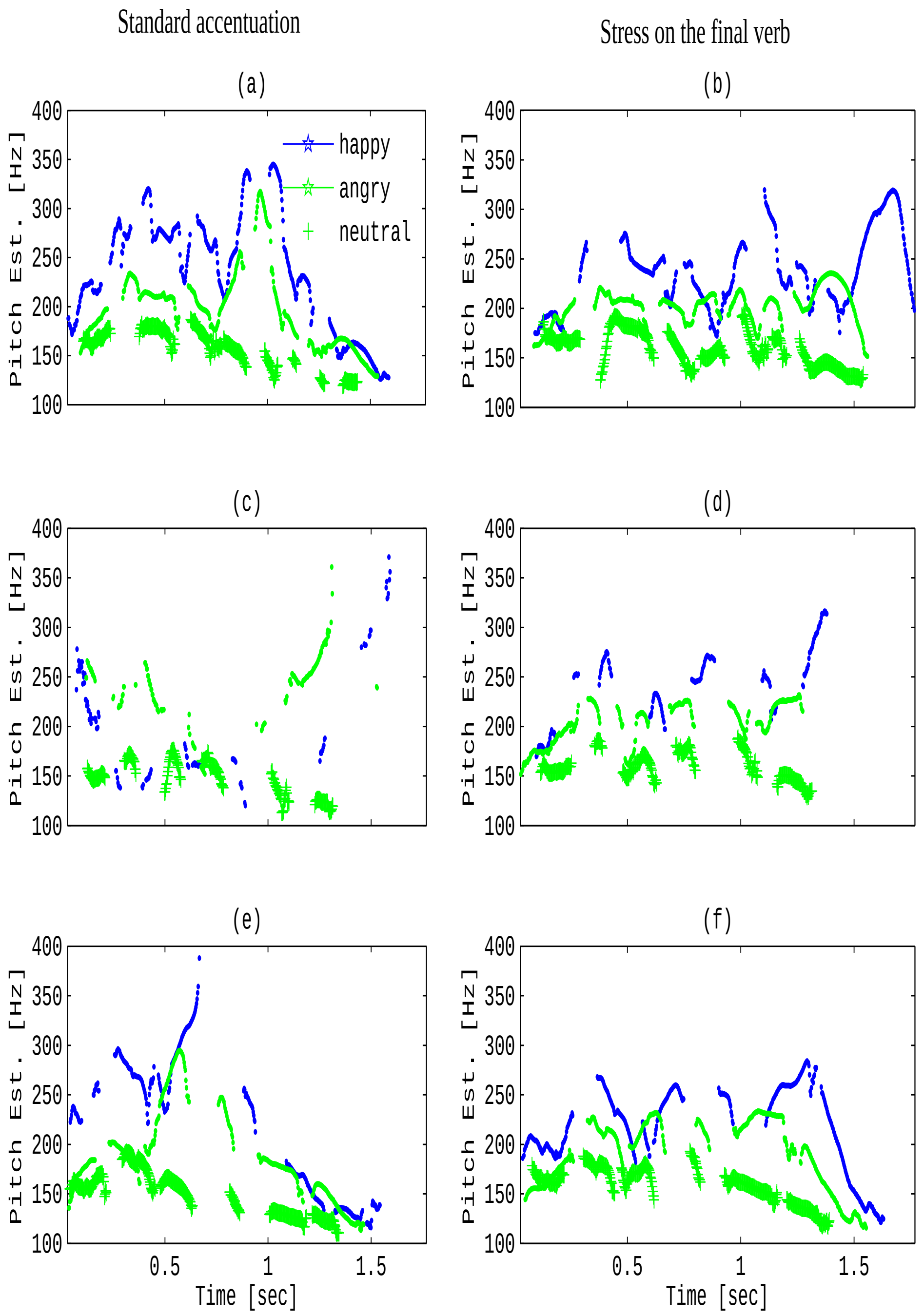

Fig. 4. F0 contours for three emotive categories as function of the two placements of stress and sentence lexical ratings. (a) Sentence $\#1(17)$ (negatively rated), stress on the noun. (b) Sentence $\#1(17)$, accent on the final verb. (c) Sentence $\#4(85)$ (positively rated), stress on the noun. (d) Sentence $\#4(85)$, stress on the final verb. (e) Sentence $\#7(123)$ (neutrally rated), stress on the noun. (f) Sentence $\#7(123)$, stress on the final verb. The legend for all plots is shown in (a) and refers to the sentence numbers used in the original corpus.

merge. The separation between F0 contours for the three emotional categories is greater in the NS case than in the VS case, as we already found from the F0 mean values.

It should also be noted that in the ANS case, sentence $\#7(123)$, the F0 contour shape is very different from that of the other utterances. In particular, it does not have its own characteristic range, but instead it hovers around the ’neutral’ NOS contour until about .4 sec into the utterance after which it rises to the level of the HAS contour. It stays within the HAS F0 range through the rest of the sentence. It appears as though the speaker was able to express the emotion through the stressed noun primarily, without the need to spend vocal effort at the beginning of the sentence. This example indicates that intonation contours can have quite different shapes. In fact, the uniformity of F0 contours for the rest of the utterances, with the few noted exceptions, may be due to the utterances being spoken by a professional actress who places maximum vocal effort into each expression. In spontaneous speech, greater variations can be expected.

To summarize, stress placement appears to have a considerable effect on the aggregated and temporal measures utilized in AAR. F0 mean, range, and utterance duration are the measures most affected by stress placement. The second part of the analysis revealed a variety of possible ’accent’ patterns in emotive speech, so that a reliable correlation of the stress placements to specific patterns could not be established. This variability suggests that it would be difficult to come up with a practical F0-based model of stress placement in emotive speech. We expect to see an even greater variability in the measures described here for spontaneous speech, not uttered by a trained speaker. The cumulative effect of all the factors examined is be expected to decrease the success of AAR.

4 Conclusion

Pitch supplies key features used in AAR. However, pitch is modulated by a multitude of other prosodic as well as non-prosodic factors. In this paper, I have examined some of the challenges in automated affect recognition that could be posed by 1) the presence of an extra prosodic variable, namely, the stress placement, and 2) by non-standard lexical content, which either was nonspecific, or at odds with the expressed affect. These questions were investigated experimentally, by examining how utterance duration and F0-based measures relevant to AAR change in the presence of these conditions.

For the set of sentences analyzed, we find that the measures, which include aggregated F0 measures - primarily mean and range - and the utterance durations, correlate with the three emotive expressions: a neutral speaking style and the two primary emotions (happiness and cold anger). However, these correlations were fundamentally affected by the placement of the stress, to the point where the F0-affect relationships were reversed in the two stress placement cases. The stress placement also influenced the *consistency* of the F0-affect relationship through the duration of the utterance. This influence manifested itself in several ways. There were differences in the amount of separation of the F0 contours (local F0 mean values) in the same part of the sentence for the two stress placements. The duration of the utterances was also influenced by stress placement, but whether the utterances became longer or shorter with the change in stress placement depended on the emotive expression of the utterance. These results suggest that the two prosodic conditions - emotion and accentuation - are not expressed independently and their combined effect on pitch is rather complex.

Examination of mid to long-range features of F0 contours did not reveal any consistent correlation between these features and stress placement. This suggests that detection of stress placement in emotive speech is difficult to perform. Influence of lexical content on F0 contours was also examined. Although an example was found where the F0 contour was distinct, the negatively rated sentence when the stress was on the verb, other possible factors could have influenced this result and so no conclusion could be drawn.

There was a single example where the shape of an F0 contour was fundamentally different from those in the rest of the examples, indicating that sizeable deviations are possible from the observed norm here. This occurred for the lexically neutral sentence uttered with angry expression and standard accentuation. The difference may be attributable to the speaker being able to express anger in a single word and therefore not needing to expend vocal resources to maintain effort throughout the sentence. While no conclusion should be drawn from a single example, this case indicates that affect expression in sentences that carry an emotional connotation may require greater vocal effort, regardless of whether the connotation is lexically matched or not to the expressed affect.

The significant influence that stress placement has on pitch in the results shown here suggests that recognition of emotions should not be considered in isolation from other prosodic factors. We note further that the magnitude of the accent did not vary in this study. If it did, we would expect to see even greater alterations in F0 measures. The addition of more prosodic variables would also likely exacerbate alterations in F0 measures. Since pitch provides essential information in AAR, such distortions could undermine the AAR recognition task.

On a final note, the utterances analyzed here were spoken by a trained actress, who, it can be assumed, was meticulous about fully articulating the designated affect and accentuation. The rate of speech was also likely slower in this case. The deliberate, heightened articulation and the slower pace of speech are likely to produce a more even and sustained vocal effort than in the case of spontaneous speech. In spontaneous speech, the lack of deliberate vocal effort may translate into a diminished and less clear differentiation among the F0 measures and could make the case regarding the implications for AAR even stronger. The same arguments can be made with regard to the observations on the lack of dependence of pitch on the lexical content. That is, the lack of differentiation seen here may be due to the actress's frank effort to provide the highest emotive expression possible in each utterance. Spontaneous speech may reveal a different picture from what is observed here.

5 Discussion

It is often suggested that machine recognition of at least a subset of key emotions would help mainstream acceptance of speech-based interfaces (SIS). However, in placing emphasis on emotions, the technology has sidestepped recognition of other prosodic cues. This paper, and a number of others, e.g. ten Bosch (2003), have raised the question whether recognition of emotions without considering other pitch- modulating elements is realistic. The question is a part of a larger issue discussed here.

Users' annoyance at a machine unable to respond to emotional cues is clearly undesirable, but it is unlikely to pose an insurmountable barrier to adoption of many speech-based interfaces, such as telephony based transactions. The true issue seems to be efficiency. Speech is slow compared to a keyboard or point-and-click devices for transmitting factual information to a machine. As such, it is inferior to visual input modalities where a complex combination of choices can be specified with a simple tap to a screen. An exception to this is presented by automated call systems, but only because phones currently are not equipped with screen menus to tap.

In addition, speech consumes limited cognitive resources, interfering with cognitive tasks. In spontaneous speech, people optimize their resources by allowing a certain degree of lexical, syntactical, and semantic imprecision. Yet, despite the prevalence of ambiguity in natural language, human-human verbal communication *is precise* in the mind of the listener, where a speakers message becomes crisp. It is suggested here that prosody plays a considerable role in disambiguating human communication. Prosody allows one to optimize the communicative content of an utterance and even reduce a full sentence to a few words. Prosodic cues can indicate the speaker's attitude, intent, the degree of importance of ongoing tasks, and levels of stress. These cues are not binary variables that could be coded into a menu. Moreover, linguistic denotation of many words, for example adjectives (large, attractive, bearable) is also a continuous and often a vague entity. Such content is not readily transferable into clicks of a mouse, and so here would seem to lie the untapped potential of speech-base interfaces. If speech technologies could handle the ambiguities in casual speech, speech-based interactive systems could provide a faster, more natural, and cognitively less demanding means of communication than hand control or visual interfaces.

The question is what defines the most important discriminants in the domain of prosodic cues? Most research in prosody has focused on spoken affect because we perceive emotions as being prevalent in our speech. This is not necessarily supported by research, i.e. Cowie and Cornelius (2003) and references therein. Perhaps what we communicate through vocal cues are more subtle information - frustration, impatience, etc. - which people subconsciously relate to emotional states. For example, low pitch is known to convey dominance, power, and confidence. A listener may subconsciously register these attributes and infer that the speaker is not afraid.

Since AAR for spontaneous speech in adverse conditions is very difficult and since data, such as that presented in this paper, suggest that spoken emotions may not be easily separated from other prosodic features, it is worth evaluating what non-linguistic information is essential for various SIS. Should we care about a user's happiness or user's level of confidence when processing a voice driven purchasing transaction? The attitude, intent, and degree of confidence or frustration are important, but many current technologies aim to infer these cues indirectly through a user's affect. A more profitable course could be to directly relate F0 measures to the user's states that are of interest.

References

Alter, K., Rank, E., Kotz, A., Toepel, U., Besson, M., Schirmer, A., Friederici, A., 2003. Affective encoding in the speech signal and in event-related brain potentials. Speech Communication 40, 61–70.

Amir, N., Ron, S., 1998. Towards an automatic classification of emotions in speech. In: Proc of the ICSLP 98.

Bnziger, T., R., S. K., 2005. The role of intonation in emotional expressions. Speech Communication 46, 252–267.

Cornelius, R., 1992. The science of emotion.

Cowie, R., Cornelius, R., 2003. Describing the emotional states that are expressed in speech. Speech Communication 40, 5–32.

Cowie, R., Douglas-Cowie, E., Romano, A., 1999. Changing emotional tone in dialogue and its prosodic correlates. In: Proc. ESCA Workshop on Dialogue and Prosody. pp. 41–6.

Dellaert, F., Polzin, T., Waibel, A., 1996. Recognizing emotion in speech. In: Proc of the ICSLP 96.

Ekman, P., 1999. Basic Emotions, in Handbook of cognition and emotion. Wiley & Sons, Ltd., NJ.

Gorodnitsky, I., 2007. Dynamical theory formalism for robust modeling of damped, undamped, and nonlinear oscillatory signals. In: Proc. IEEE Int. Conf. on Acoustic, Speech and Sig. Processing. Vol. III. pp. 725–28.

Heuft, B., Portele, T., M., R., 1996. Emotions in time domain synthesis. In: Proc of the ICSLP 96.

Iida, A., Campbell, W. N., Iga, S., Higuchi, F., Yasumura, M., 1998. Acoustic nature and perceptual testing of corpora of emotional speech. In: Proc of the ICSLP 98.

Lang, P., 1995. The emotion probe. studies of motivation and attention. Am. Psychologist 50 (5), 372–85.

Li, Y., Zhao, Y., 1998. Recognizing emotions in speech using short-term and long-term features. In: Proc Int Conf on Speech and Language Processing. pp. 2255–58.

Mozziconacci, S., Hermes, D., 1999. Role of intonation patterns in conveying emotion in speech.

Noad, J. E. H., Whiteside, S. P., Green, P. D., 1997. A macroscopic analysis of an emotional speech corpus. In: Proc EuroSpeech ’97.

Noam, A., 2001. Classifying emotions in speech: A comparison of methods. In: EuroSpeech ’01.

Noquerias, A., Moreno, A., Bonafonte, A., Marino, J., 2001. Speech emotion recognition using hidden markov models. In: EuroSpeech ’01.

Norman, D., 2004. Emotional Design: Why We Love (Or Hate) Everyday Things. Basic Books, New York.

Picard, R., 1997. Affective Computing. MIT Press, Cambridge.

Sjlander, K., 2005. The snack sound visualization module, version 2.2.10. http://www.speech.kth.se/snack.

ten Bosch, L., 2003. Emotions, speech and the asr framework. Speech Communication 40, 213–225.