Prosodic cues for emotion characterization in real-life spoken dialogs[1]

Laurence Devillers ◊ *and Ioana Vasilescu*♠

◊ LIMSI-CNRS 0rsay, France
♠ ENST-CNRS, TSI, Paris, France

devil@limsi.fr, vasilesc@tsi.enst.fr

Abstract

This paper reports on an analysis of prosodic cues for emotion characterization in 100 natural spoken dialogs recorded at a telephone customer service center. The corpus annotated with task-dependent emotion tags which were validated by a perceptual test. Two F0 range parameters, one at the sentence level and the other at the subsegment level, emerge as the most salient cues for emotion classification. These parameters can differentiate between negative emotion (*irritation/anger, anxiety/fear)* and *neutral* attitude and confirm trends illustrated by the perceptual experiment.

1. Introduction

In recent years there has been growing interest in the study of emotions [1, 5, 9] to improve the capabilities of current speech technologies (speech synthesis, speech recognition, and dialog systems). In the context of human-machine interaction, the study of emotion has generally been aimed at the automatic extraction of mood features in order to be able to dynamically adapt the dialog strategy of the automatic system or for the more critical phases, to pass the communication over to a human operator.

According to Scherer [12] the first problem in analyzing emotions is the difficulty of isolating the emotion factors, as it is closely related to several other human behaviors, such as mood, interpersonal stances, attitudes, and personality traits. The second reason of complexity is related to the fact that “full blown”, pure, basic emotions do not occur frequently in spontaneous verbal interactions. Instead are identified in such interactions mostly shaded, mixed, blended etc., emotions which are difficult to isolate, describe and detect. Most of the studies have only focused on a minimal set of emotions or attitudes such as four primary emotions (*anger, fear, sadness* and *joy*), positive/negative emotions [11], emotional/neutral state [1] or stressed/non stressed speech [8].

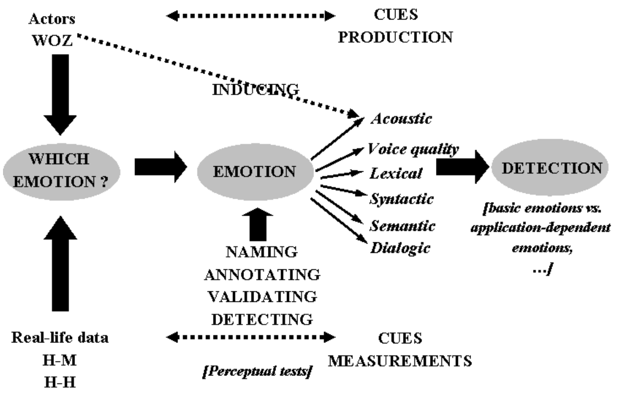

Figure 1: *Emotion production, analysis and detection*

Emotion manifestations are strongly dependent on the corpus employed. Figure 1 synthesize the current methods in emotion production, analysis and detection. More precisely, three types of corpora are typical studied: actors, WOz and real-life data. The corpora obtained by asking actors to simulate emotions is the easiest to exploit, as the semantic and lexical levels are controlled and the emotions markers are mainly expressed at prosodic level. For this type of data, emotion detection is possible with just acoustic information. However, the closer we get to the real-life context of interaction, the more difficult the detection of reliable emotion markers will be and there is increasingly strong evidence that results based on laboratory research with archetypal states transfer poorly to real applications.

In [1] several levels of emotion carrying information, (i.e., prosodic, parts of speech, dialog acts, syntactic-prosodic boundaries, repetitions etc.) were employed for emotion detection in a WOz corpus. Similar work using both human-human and human-computer dialogs focused on emotion detection with a mixture of traditional acoustic and linguistic information [9]. However, real-life corpora are poorly represented in the literature despite the general agreement that the research should encourage a natural (real-life) point of view in collecting databases, i.e., highlighting everyday expressions of emotion: news, spontaneous conversations, call centers, medical emergency services etc.

[1] This work was partially financed by the European Commission under the IST-2000-25033 AMITIES project http://www.dcs.shef.ac.uk/nlp/amities.

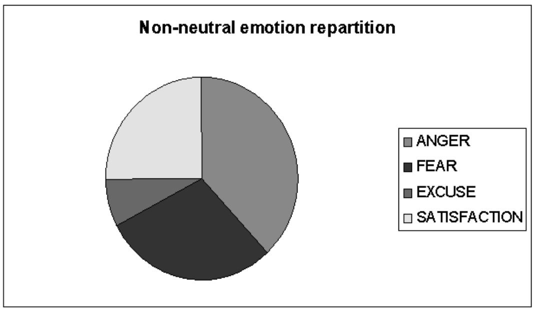

Figure 2: *Proportion of the non-neutral emotion labels (13% of the corpus)*

The present study is carried out within the framework of the IST Amities *(Automated Multi-lingual Interaction with Information and Services)* project, and makes use of a corpus of real agent-client dialogs recorded in French (for independent purposes) at a Stock Exchange Customer Service Center.

Previously, we have reported on evaluating the role of lexical and dialogic information respectively in emotion detection [6]. The final aim of our research is to build a multi-level detection model in which prosodic, lexical and dialogic levels contribute to the final detection score.

In the following sections, we present the results obtained on the prosodic level. In section 2, we describe the corpus and the annotations. Section 3 gives trends in perceived prosodic cues. Section 4 reports results on the prosodic cues measurements on the global corpus and per dialog. Conclusions and further research are discussed in section 5.

2. Corpus and annotation

Our corpus consists of about 5000 speaker turns (100 clients, 4 agents) extracted from 100 dialogs. The corpus covers a large range of possible spontaneous realizations in terms of topics, sentence lengths and types (interrogative, assertive etc.) and speaker characteristics (voice quality, gender etc.).

Two annotators independently listened to the 100 dialogs, labeling each sentence (agent and customer) with one of the five emotions (*anger, fear, satisfaction, excuse, neutral attitude*). Sentences with ambiguous labels ($\sim$ 3%) for those annotations were judged by a third independent annotator. Around 13% of the corpus (660 sentences) when the audio and dialogic context are available, are annotated with non-neutral emotion. Figure 2 gives the proportion of the different non-neutral emotion labels.

Systematic and careful evaluations of emotion tagsets are generally lacking. In order to validate our annotations, we conducted a perceptual test on 40 sentences representing the 5 emotion classes as material. The experimental protocol is described in [7]. The test consisted of naming the emotion present in each stimulus and of describing the prosodic cues. These results validate the presence of emotion manifestations. In fact, *anger* is perceived as *irritation* and *fear* as *anxiety*. Therefore, we replace basic emotions with shaded emotion marks *anger* → *irritation* and *fear* → *anxiety*. An interesting result is that *satisfaction* is globally perceived as *neutral* by subjects. We can explain this finding by satisfaction marks which generally indicate a normal dialog progression. The perceived prosodic cues are discussed further in the next section.

	Perceived prosodic cues	
	Value	Emotion
Speaking Rate	Slow	-
	Normal	Anxiety
		Neutral
	Fast	**Irritation**
		Satisfaction
		Excuse
δF0	**Flat**	**Neutral**
		Excuse
	Variable	Other emotions
Energy	Normal, Low, High	All emotions

Table 1: *Perceptual classification of main prosodic features.*

3. Perceived prosodic cues

The classical prosodic parameters associated with emotions are: speech rate, F0 variation and energy. In addition, other acoustic factors may contribute to vocal emotion detection: formants and temporal features such as pausing, hesitation, segment lengthening.

In [7], we focused on parameters allowing a perceptual "naive" description. We asked the subjects to mention their perceptual feeling on the melody (F0 variation), speed (speech rate) and energy (loudness) after listening to the speech signal.

The choices proposed for the speech rate were: slow, normal and fast; for energy: normal and high; and for F0 variation: flat or variable. The majority of subjects (see Table 1) judged the speech rate as fast for *irritation* and *satisfaction*, whereas the F0 variation allowed subjects to distinguish *neutral* state and *excuse* (flat) from other emotional states (variable). We did not observe any systematic energy variation that could be related to emotion.

In contrast to the perceptual cues found to be relevant using simulated emotions produced by actors (which are often expressed with more prosodic clues than in realistic speech data), in WOz and real-life corpora the prosodic cues are much less easily identifiable as callers may use multiple linguistic strategies. The perceptual test revealed the melody (F0 variation) and the speed (speech rate) as the main prosodic perceived cues. Furthermore others studies point to the F0 as the main prosodic cue

F0 variation (sentence level)					
Labels	*Ang*	*Fea*	*Sat*	*Neu*	*Exc*
range F0 (Hz)	++	+	=	=	−
max δF0 (Hz)	++	=	=	=	−

Table 2: *Trends (Mean Values) for emotion effects on selected prosodic parameters correlated with emotion classes for perceptual test subset (40 speaker turns). Symbols: ++: very high, +: high, =: medium, -: low. Ang= anger/irritation, Fea=fear/anxiety, Sat=satisfaction, Neu=neutral, Exc=excuse.*

for emotion detection. This work focuses on the F0 features, as described next.

4. F0 features

We have used the PRAAT program [2] to extract F0 features (measures estimated for the pitch) and voiced segments. It is based on a robust algorithm for periodicity detection, working in the lag (auto-correlation) domain. This algorithm is particularly adapted for noise condition (telephone speech) and allows to detect specific acoustic phenomena. Among the vocal manifestations of emotion we have noticed that a rapid change in voice quality is a way to express negative emotions.

In this study, we estimated the classical F0 measures (min, max, mean, range, standard deviation) for each speaker turn (sentence level). For F0 calculation, only voiced regions were taken into account. We also calculated the maximum cross-variation of F0 between two adjoining voiced segments δF0 (sub-segment level). The high values obtained on short segments ($<$ 40 ms) have been considered detection errors and thus eliminated.

The F0 parameters were computed for the entire corpus. An analysis is given for the perceptual subset and the full corpus. The F0 parameters (min, max, mean, standard deviation) emerge poorly as related to emotion. Our study focuses on the more relevant parameters which are range F0 and max δF0. The range is measuring F0 variation without any information of the distribution of F0 values within that range. The max δF0 parameter is a measure of local variation. It can capture rapidly changes in voice quality.

4.1. Perceptual test subset analysis

For the 40 sentences of the perceptual subset, the two parameters: range F0 and max δF0 are the most salient for emotion classification (see Table 2). Given the small amount of data, we mention trends of mean values. According to the two parameters, three emotion groups emerge *anger/irritation*, *fear/anxiety* and (*neutral, satisfaction* and *excuse*) (Table 2). Among negative emotions, *fear/anxiety* is less expressed on F0 variation. These results are correlated with the perceived cues on negative emotions.

F0 variation (sentence level)					
Labels	*Ang*	*Fea*	*Exc*	*Sat*	*Neu*
number of sentences	253	192	51	167	4295
range F0 (Hz)	**220**	**228**	201	174	171
max δF0 (Hz)	**129**	**127**	97	91	81

Table 3: *Mean values for emotion effects on selected prosodic parameters correlated with the 5 emotions on the full corpus (5K speaker turns). Symbols: Ang= anger/irritation, Fea=fear/anxiety, Sat=satisfaction, Neu=neutral, Exc=excuse.*

4.2. Full corpus analysis

The acoustic variability in the full corpus can be attributed to two main factors: environmental conditions and intra- and inter-speaker variability. The different speakers have their own individual voice characteristics and acoustic correlates of emotions which are particularly interesting for our study. Therefore, we analyzed the two selected parameters (range F0 and max δF0) from two points of view: sentence-level and dialog-level.

4.2.1. Sentence-level

Comparing for instance these simple parameters allows to notice a strong difference between two groups of values: the group of negative emotion values *anger/irritation, fear/anxiety* have higher measures than the others (see Table 4), thus strengthening the trends provided by the subset of corpus used in the perceptual tests. This difference is statistically significant (t-test $p < 0.001$) as shown by a parametric comparison of the range and δF0 values for negative versus neutral emotions. The neutral group of emotions contains *neutral* and also *excuse* and *satisfaction*. The *excuse* marks have intermediate values at least for range F0. The excuse are mainly expressed at lexical level. *Satisfaction* marks are perceived like very closed to *neutral* attitude in this particular application and confirm the perceptual cues indicated by the subjects for this class. Given the large speaker variability of the corpus, results show only general trends and need to be verified at dialog level.

4.2.2. Dialog level

We carried out analysis for each dialog as most relevant for our types of applications (call center). We have calculated the F0 parameters (range and δF0) for the client sentences and for each dialog. This experiment gives a more realistic measure of the parameters saliency by taking into account each speaker's variability.

From the 100 dialogs of the corpus, 76 dialogs were annotated with both negative and neutral client emotional manifestation. Three ratios were calculated at the dialog level; R1: the percentage of dialogs in which both F0 range parameters for the negative *fear/anxiety,*

Range F0 and δF0 cues (Speakers variability)	
Ratios	*% of speakers*
R1: cues(Fea) & cues(Ang) > cues(Neu)	61%
R2: cues(Ang) > cues(Neu)	75%
R3: cues(Fea) > cues(Neu)	68%

Table 4: *Speakers following the trends found with the selected prosodic parameters for emotion effects. The selected cues are "Range F0" and "δF0".*

anger/irritation emotions are superior to *neutral*, R2: the percentage of dialogs in which both F0 range parameters for *anger* are superior to *neutral* and R3: the percentage of dialogs in which both F0 range parameters for *fear/anxiety* are superior to *neutral*. The results are given in Table 5. They show that 46 of the 76 dialogs (61%) have a difference between negative emotions cues versus neutral emotions cues underlying the trends found on the corpus. Regarding each negative emotion separately, we observe that the analyzed prosodic cues are more useful for *anger/irritation* then for *fear/anxiety*. Among the remaining subset of dialogs (24 dialogs), 70% show only *neutral* emotions with lower parameters (range F0 and max δF0) than those of the sentences labelized with negative emotions. The other 30% of the dialogs show both *satisfaction* and *neutral* emotion marks. For three of these last dialogs, F0 range parameter for *satisfaction* have the same magnitude as for negative.

In order to illustrate the different strategies in expressing emotions, we selected 2 dialogs in which two different clients experience *anger/irritation* and *fear/anxiety* (Table 5). The first dialog provides values for the two F0 parameters which follow fully the trends for negative vs *neutral* emotion (R1). The second dialog confirms the trends uniquely for prosodic cues for *anger/irritation* (R2), and the marks for *fear/anxiety* are particularly poor. However the sentences corresponding to those marks have been clearly annotated as *fear/anxiety* suggesting that the emotion marks belong to another level. *Fear/anxiety* is often associated with repetitions and disfluencies [6]. These results suggest that the prosody is not the only strategy to express emotion, and the lexical and dialogic cues have been employed as well.

Examples of FO variation for 2 dialogs					
Client 1 (woman)					
Labels	*Ang*	*Fea*	*Exc*	*Sat*	*Neu*
number of sentences	2	3	-	5	13
range F0 (Hz)	224	340	-	134	190
max δF0 (Hz)	211	152	-	41	75
Client 2 (man)					
Labels	*Ang*	*Fea*	*Exc*	*Sat*	*Neu*
number of sentences	10	2	-	-	20
range F0 (Hz)	292	73	-	-	206
max δF0 (Hz)	170	26	-		120

Table 5: *Mean values for all sentences and for emotion effects on selected prosodic parameters correlated with the 5 emotions for two speakers.*

5. Conclusion

In this study we describe some recent experiments to locate acoustic features which indicate the presence of non-neutral emotions in a dialog corpus. Two parameters emerged from the analysis : the F0 range and the max δF0. These parameters validate trends found in perceptual experiments and can serve to classify emotions in a client-agent corpus recorded in a call center. We observe a strong correlation between the two parameters and negative (*fear/anxiety, anger/irritation*) versus *neutral* emotion. These prosodic features are correlated with trends found in the literature.

Our ongoing work focuses on automatically determining the F0 measures at the syllable level in order to correlate the prosodic cues (F0, pause) with lexical and semantic information. In addition, other prosodic cues (F0 slope, segment lengthening, hesitation) are also being studied.

As emotions in real-life interaction have complex manifestations integrating several linguistic levels and/or non linguistic markers, our future work will explore the combination of emotion information conveyed by the lexical, semantic and contextual information with prosodic features.

6. References

[1] A. Batliner et al., "How to find trouble in communication", *Speech Communication*, 2003.

[2] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", '*IFA Proceedings*, 1993, p 97-110.

[3] N. Campbell, " Recording techniques for capturing natural everyday speech" *LREC*, Las Palmas, 2002.

[4] E. Douglas-Cowie, N. Campbell, R. Cowie and P. Roach, "Emotional speech; Towards a new generation of databases", *Speech Communication*, 2003.

[5] F. Dellaert, T. Polzin, A. Waibel, "Recognizing Emotion In Speech," *ICSLP*, 1996.

[6] L. Devillers, I. Vasilescu, L. Lamel, "Emotion detection in task-oriented dialogs corpus", *ICME*, Batimore, July 2003.

[7] L. Devillers, I. Vasilescu, C. Mathon, "Prosodic cues for perceptual emotion detection in task-oriented Human-Hum an corpus",*ICPhs*, Barcelone, August 2003.

[8] R. Fernandez, R. Picard, "Modeling Drivers' Speech Under Stress," *Speech Communication*, 2003.

[9] C.M. Lee, S. Narayanan, R. Pieraccini, "Recognition of Negative Emotions from the Speech Signal",*ASRU*, 2001.

[10] S. Narayanan, "Towards modeling user behavior in human-machine interactions: Effect of Errors and Emotions" , *ISLE Worshop*, Edinburgh 2002.

[11] C.M. Lee et al., "Combining acoustic and language information for emotion recognition", *ICSLP*, 2002.

[12] K. Sherer, "Vocal communication of emotion: A review of research paradigms" *Speech Communication*, 2003.