



F0 and pause features analysis for Anger and Fear detection in real-life spoken dialogs

L. Devillers[◇] & I. Vasilescu[♣] & L. Vidrascu[◇]

[◇] LIMSI-CNRS Orsay, France,
[♣] ENST-CNRS, TSI, Paris, France

devil@limsi.fr, vasilescu@tsi.enst.fr, vidrascu@limsi.fr

Abstract

This paper describes recent work focusing on F0 and pause features detection for two negative emotions, *Anger* and *Fear*, occurring in real-life human-human spoken dialogs. Most of the current studies do not differentiate within the class of negative emotions, when an automatic system should consider appropriate strategies according to different negative emotions. In this paper we consider two types of prosodic cues aiming to differentiate between two negative emotions *Anger* and *Fear*. The work is carried out in the context of the AMITIES project in which spoken dialog systems for call center services are being developed. F0 features are two range parameters, one at the sentence level and the other at the sub-segment level. Pause features are meaningful silent pauses and filler pause “euh”. We correlate all the features with emotion labels and with two variables, gender and speaker (agent vs client). The study shows that pause features are a global more reliable cue to distinguish between *Anger* and *Fear* than F0 parameters. However, differences in both F0 and pause patterns needs to be made according to speaker and dialogic context.

1. Introduction

In recent years there has been a growing interest in the study of emotions [1][3][11] in order to improve the capacities 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.

Most of the studies focus on the opposition negative/positive emotions. However distinctions should also be made inside the negative class. According to the type of negative emotion, the system will adopt a different strategy. The question we address here is whether the two main negative emotions *Anger* and *Fear* present in our corpus show prosodic manifestations robust enough to differentiate them.

According to Sherer, the literature on emotions define *Anger* as being vocally expressed by an increase in mean F0 and mean intensity as well as in F0 variability manifested in increased F0 range. Further *Anger* signs seem to be an increase in high frequency energy and downward directed F0 contours. The rate of articulation increases as well. Concerning *Fear*, the data shows increases in mean F0, F0 range and in high frequency energy. Rate of articulation seems to increase as well

[14]. It appears that the two emotions have quite similar manifestations. For other researchers, manifestations for *Fear* have more intense F0 patterns than for *Anger* [16]. As Scherer [15] has pointed out, there is an apparent contradiction between the difficulty in finding acoustic differentiation of emotional states and the comparative ease with which listeners are able to judge emotions from speech.

In previous studies on the Amities corpus [2], we have shown that F0 range variations allow to distinguish between negative and positive emotions. Finally, we have found that during perceptual tests subjects are able to differentiate *Anger* and *Fear* [9] with 75% accuracy.

In this paper we aim to analyse F0 and pause features enabling to differentiate between *Anger* and *Fear*. We correlate F0 and pauses features with the emotions labels given two variables: gender (male/female) and speaker (agent/client).

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. In the following sections, we describe the corpus and the data processing (section 2) and the analysis of the F0 and pause features (section 3). Conclusions and further research are discussed in section 4.

2. Corpus

The dialogs are real agent-client recordings from a Web-based Stock Exchange Customer Service center. These recordings were made for purposes independent of this study, and have been made available for use in developing an automated call routing service within the context of the AMITIES project. The service center can be reached via an Internet connection or by directly calling an agent. While many of the calls involve problems in using the Web to carry out transactions (general information, complicated requests, transactions, confirmations, connection failures), some of the callers simply seem to prefer interacting with a human agent. A corpus of 100 agent-client dialogs (4 different agents) in French has been orthographically

Table 1: Characteristics of the corpus of 100 agent-client dialogs.

# agents	4	# clients	100
# turns/dialog	ave: 50	min: 5	max: 227
# words/turn	ave: 9	min: 1	max: 128
# words total	44.1k	# distinct	3k

Table 2: Proportion of each emotion label in the dialog corpus labeled by listening to the audio signal.

	<i>Anger</i>	<i>Fear</i>	<i>Satisfaction</i>	<i>Excuse</i>	<i>Neutral</i>
<i>Client</i>	9.9%	6.7%	2.6%	0.1%	80.7%
<i>Agent</i>	0.7%	1.3%	4.0%	1.8%	92.1%

transcribed and annotated. The dialogs cover a range of investment related topics such as information requests (services, commission fees, stock quotations), orders (buy, sell, status), account management (open, close, transfer, credit, debit) and Web questions and problems. Table 1 summarizes the characteristics of the corpus. There are 6241 speaker turns. We considered 5000 sentences after excluding overlaps which are known to be frequent phenomena in spontaneous speech.

3. Speech data and processing

3.1. Emotion annotation

A task-dependent annotation scheme was developed, keeping in mind that the basic affective disposition towards a computer is generally either trust or irritation. Three of the five classical emotions are retained: *Anger* (A), *Fear* (F) and *Neutral* (N) attitude (the normal progression of the dialog). In this Web-based stock exchange context, most of *Anger* and *Fear* manifestations are shaded emotions such as nervousity or irritation for *Anger*, and worry or anxiety for *Fear*.

We also considered some of the agents’s and client’s behaviors directly associated with the task in order to capture some of the dialog dynamics. For this purpose, *Satisfaction* (S) and *Excuse* (E) (apology) were included in the emotion labels. These correspond to a particular class of the speech acts as described in the classical version of pragmatic theory.

Two annotators independently listened to the 100 dialogs, labeling each of the sentences with one of the 5 emotions. In order to assess the consistency of the selected labels, we conducted perceptive tests and calculated the inter-annotation agreement. Ambiguities concern 2.7% of the corpus and most often involved indecision between neutral state and other emotion.

Sentences with ambiguous labels (19% of the sentences labeled with non-neutral emotion labels) were judged by a third independent listener in order to decide on the final label.

Based on the auditory classification, sentences with non-neutral labels (F, A, S, E) comprise about 13.2% of the entire corpus. The proportion of non-neutral emotions *Anger* and *Fear* for clients is 8 times (2.10% for agents vs 16.5% for clients) higher than for agent. More precisely, among the utterances labeled *Anger*, 7.5% belong to the agents and 92.5% to the clients, whereas the ratio for *Fear* is 20% vs 80%. *Excuse* characterizes agents’ turns, *Satisfaction* is twice more frequent in agents’ sentences and *Neutral* is equally distributed between agents and clients.

3.2. F0 processing and normalization

PRAAT has been used to extract F0 features on voiced regions. 1.4% of short segments (< 40 ms) have been considered detection errors and eliminated. These errors are homogeneously distributed among all the 5 classes. Two F0 measures are considered for each speaker turn: at the (*sentence level*) the F0 range

range F0 and at the sub-segment level the maximum cross-variation of F0 between two adjoining voiced segments **max $\delta F0$** (*sub-segment level*). They illustrate extreme manifestation in F0 variations.

The z-score normalisation method has been used. It is computed by removing the mean obtained over all values of a speaker in a dialog and dividing by the corresponding standard deviation.

3.3. Automatic alignment for pauses extraction

We proceed to an automatic alignment of the orthographic transcription with the acoustic signal in order to extract additional prosodic cues. The orthographic transcriptions are aligned with the signal using existing models already developed at LIMSI for another task (telephonic conversations) [?]. The alignment system uses continuous density HMMs with Gaussian mixture for acoustic modeling. The vocabulary contains 3022 words with a phonetic transcription based on 37 phones. Each context-dependent phone model is tied-state left-to-right CD-HMM with Gaussian mixture observation densities (16 per state). 4.6% of utterances have not been automatically aligned. Among them, 40 utterances corresponding to negative emotions have been manually aligned. In addition, all the utterances labeled with negative emotions have been manually verified in order to avoid alignment errors. In this study we have not considered *Excuse* and *Satisfaction* utterances and the negative emotions have only been compared with neutral state. The new pause features relying on the automatic alignment information are silent pauses and filler pause "euh".

4. F0 and pause features

The most salient parameters (range F0 and $\delta F0$) have been correlated with main emotion classes (negative/positive) at two levels of analysis: sentence-level and dialog-level [2]. In this study we differentiate between agent/client and male/female emotion manifestations.

4.1. F0 analysis for agents

Concerning the negative class of emotions, F0 inter-agent variations show three different strategies (see Table 4.1). Two male agents (agent 1+2) show similar strategies in which *Fear* is poorly represented and the only negative emotion strongly represented is *Anger*. Furthermore, manifestations for *Fear* are less acute than for *Neutral* class, however we have to consider than the number of utterances for *Neutral* is more important and better illustrate the class. One male agent (agent 3) shows stronger negative emotional behavior as both parameters are higher than for the others agents. Finally, the woman agent (agent 4) shows less F0 variation for both parameters when angry, but she has more variation when she is experiencing *Fear*. Both agent 3 and 4 show higher values for negative emotions than for the *Neutral* class. Those results highlight the idea that emotions have extremely variable manifestations and are speaker and conversational context dependent.

4.2. F0 analysis for clients

We analyze the F0 variation results for clients, globally and according to gender. The main observation is an higher F0 global variation for *Fear* than for *Anger* for clients whereas agents show the opposite trends. This observation concerns mainly the male speakers. As a general trends, manifestations for negative

Table 3: Mean values for emotion effects on selected prosodic parameters according to gender correlated with the 3 emotions (5000 speaker turns). Symbols: Ang= Anger/Irritation, Fea=Fear/Anxiety, Neu=Neutral.

F0 inter-agent variation			
Labels	Ang	Fea	Neu
agent1 (male) - range F0 (Hz)	207	87	117
agent1 (male) - max $\delta F0$ (Hz)	111	43	60
agent2 (male) - range F0 (Hz)	122	65	102
agent2 (male) - max $\delta F0$ (Hz)	76	22	50
agent3 (male) - range F0 (Hz)	141	166	121
agent3 (male) - max $\delta F0$ (Hz)	96	104	56
agent4 (female) - range F0 (Hz)	132	193	125
agent4 (female) - max $\delta F0$ (Hz)	95	105	56

Table 4: Mean values for emotion effects on selected prosodic parameters correlated with the 3 emotions (5000 speaker turns). Symbols: Ang= Anger/Irritation, Fea=Fear/Anxiety, Neu=Neutral

F0 variation (client)			
Labels	Ang	Fea	Neu
number of sentences	234	158	1911
range F0 (Hz)	237	249	190
max $\delta F0$ (Hz)	130	137	82

emotions show a higher magnitude than for *Neutral*. Female speakers show more moderate variations for both emotion class but the global pattern is respected, i.e. a higher variation for *Fear*. As for agents, the results point out that the male manifestations for negative emotions are more acute than for female group. However, we have to keep in mind that the two classes according to gender are not balanced, i.e. 91 male speakers vs 9 female speakers.

As a last observation, we can notice that F0 values for both cues are higher for client than for agent. (see Table 4.3)

4.3. Agents' vs clients' F0 variations for negative emotions

We calculated F0 variations for clients given the three classes of agents (agent 1+2, agent 3, agent 4) described above. It appears that agents' and clients' emotional behaviors are inter-correlated (see Table 4.3). More precisely, the correlation depends on the agent emotional profile and on the dialog/situation context. Thus, agents 1+2 show similar F0 variations, i.e. more variation for *Anger* than for *Fear*. Their clients have an opposite behavior, i.e. more variation for *Fear* than for *Anger*. Male agent 3 shows extreme emotional behavior compared to his colleagues, i.e. acute F0 variations for *Anger* and *Fear*. His behavior influences the clients' attitude and elicits similar manifestations. Accordingly, when interacting with agent 3, clients show higher F0 variation for both negative emotions. Finally, agent 4 shows low variation for *Anger* and higher variation for *Fear*. Her clients experience opposite emotion manifestations. This analysis allows to hypothesize that emotion manifestation is complex and depends on the topic of the dialog (i.e. the reason of the call) but also on each

Table 5: Mean values for emotion effects on selected two F0 parameters (Range and Max $\delta F0$) correlated with the negative emotions (5000 speaker turns). Symbols: Ang= Anger/Irritation, Fea=Fear/Anxiety

Clients F0 variation for Fear and Anger given the agent		
agent	Ang	Fea
agent 1+2	238	244
	119	138
agent 3	242	279
	150	150
agent 4	222	231
	139	113

Table 6: Mean values for emotion effects on selected prosodic parameters according to gender correlated with the 3 emotions (5K speaker turns). Symbols: Ang= Anger/Irritation, Fea=Fear/Anxiety, Neu=Neutral.

F0 variation with gender			
Labels	Ang	Fea	Neu
male (91 clients) - range F0 (Hz)	240	252	180
male (91 clients) - max $\delta F0$ (Hz)	130	138	83
female (9 clients) - range F0 (Hz)	196	227	112
female (9 clients) - max $\delta F0$ (Hz)	124	125	72

dialog management by agents and clients. Moreover, the respective F0 variations for agents and clients are complex and inter-dependent.

5. Pause features

In this paragraph we analyze two pause features: silent pauses and filler pause "euh".

5.1. Silent pauses

Assuming that negative emotions allow to produce unexpected speech break more than neutral behavior, we calculated the mean number of silences and the mean for the maximum duration of silences per utterance and emotion class. We differentiate between meaningful silent pauses, aiming to find a continuation of an utterance vs non linguistic pauses, i.e. occurring when the speaker is searching an information on the internet (related to the specificity of the corpus). We considered as meaningful silences inter lexical silences between 150 and 800 ms. After 800 ms we observed that silences are mainly related to different causes than the linguistic content of the utterance, such as internet problem encountered by agent/client or waiting time during which the agent is searching information elsewhere (colleagues etc.). Those two measures are correlated with negative emotions *Anger* and *Fear*. The Table 5.1 shows that silences are differently employed by agents and clients and also according with the two emotions. The main observation is that for both agents and clients, silences are longer when occurring in an utterance labeled *Fear*. We can hypothesize that *Fear* is more subject to speech breaks than *Anger*. Generally speaking, silences occur more frequently in the utterances produced by clients but inside the negative class of emotions, differences can be noticed

Table 7: Mean values for the silent and filler pauses parameters. Symbols: *Emot-Sp*= Emotion-Speaker, *Nbutt*= Number of utterances, *Nbsil*=number of silences * 100/duration of sentence in cs, *Maxsil*= max silence per sentence, *Nbhes*= Number of filler pauses * 100 / sentence duration in cs, *Maxhes*= max filler pause per sentence

Pauses features					
<i>Emot-Sp</i>	<i>Nbutt</i>	<i>Nbsil</i>	<i>Maxsil</i>	<i>Nbhes</i>	<i>Maxhes</i>
<i>Fear-agent</i>	31	11	100	2	1200
<i>Fear-client</i>	148	9	151	5	2300
<i>Anger-agent</i>	19	3	55	2	7500
<i>Anger-client</i>	242	7	114	3	1900

depending on speakers specificities. Silent pauses are also significantly longer for both emotions when produced by clients. Thus, the role of the agent could avoid long silences.

5.2. Filler pause "euh"

We considered the autonomous main French filler pause "euh". It occurs as independent item and it has to be differentiated from vocalic lengthening. We correlate the filler pause with emotions. This correlation follows the orthographic (lexical) transcription of the dialogs and consider the number of occurrences of transcribed "euh" per emotion class. "Euh" can be correlated mainly with Fear 127 sentences (4.2%), followed by Anger 101 sentences (3.1%) and finally the other emotions. After the automatic alignment, we calculated the mean for the maximum duration of filler pause and the mean number of filler pause per utterance and emotion class. The results are shown in Table 5.1. As for the percentage of "euh" per emotion class obtained with orthographic transcription, the two values follow the trends obtained with the silent pauses. Thus, clients are more hesitating than agents: their hesitations occur more often but they are not significantly longer as the results for *Anger* show. However, filler pauses happen more frequently in the utterances labeled *Fear*.

6. Conclusion

In this paper we aimed to differentiate between two negative emotions by using F0 variation parameters and pause features. We considered F0 variation according to several variables encountered in our corpus, i.e. speaker (agent/client), gender and inter-speaker relationship. F0 variations do not allow to globally differentiate between *Anger* and *Fear*, especially for clients. Indeed, a global parameters estimation hides an intern variation depending on the role of the speaker in the dialog (agent or client) and on the more general differences due to the gender. However, F0 variation show reliable patterns when considering different classes of speakers: agents and clients, male and female and finally clients given the agent. The division of speakers in classes allows to observe the following patterns: opposite variation for *Anger* and *Fear* for clients (*Anger_i*/*Fear_i*) and agents (*Anger_i*/*Fear_i* - globally and at least for agent 1, 2 and 3); globally higher variation for male than female speakers; for both agent and client female speakers values for *Fear* are higher than values for *Anger* but the reduced number of female speakers does not allow to generalize the observation. We can conclude that F0 variation show different tendencies and highlight contextual and speaker dependent manifestation for

negative emotions but does not represent an unique cue for general distinction inside the class of negative emotions. Pause features provide new elements for *Anger* vs *Fear* differentiation. Thus, both silent and filler pause show a higher correlation with *Fear* than with *Anger*. However, when considering the variable speaker (agent vs client) this general pattern is not entirely respected. We can conclude on pause features that as for F0 variation, silent and filler pauses represent a reliable cue to improve the model of negative emotions, but inside the class of negative emotions the results are speaker dependent. Thus, the dialogic context need to be considered when analyzing the prosodic parameters. Further work will focus on analyzing the intralexical filler pauses (i.e. vocal lengthening) and speech rate in order to complexify our model.

7. References

- [1] B. Wreded, E. Shriberg, 2003, "Spotting "Hots Spots" in Meetings: Human Judgments and Prosodic Cues"; *Eurospeech*.
- [2] L.Devillers, I. Vasilescu, 2003, "Prosodic cues for emotion characterization in real-life spoken dialogs"; *Eurospeech*.
- [3] A. Batliner et al., 2003, "How to find trouble in communication", *Speech Communication*.
- [4] P. Boersma, 1993, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", *IFA Proceedings*, 97-110.
- [1] N. Campbell, 2002, "Recording techniques for capturing natural everyday speech" *LREC*, Las Palmas.
- [6] E. Douglas-Cowie, N. Campbell, R. Cowie and P. Roach, 2003, "Emotional speech; Towards a new generation of databases", *Speech Communication*.
- [7] F. Dellaert, T. Polzin, A. Waibel, 1996, "Recognizing Emotion In Speech," *ICSLP*.
- [8] L. Devillers, I. Vasilescu, L. Lamel, 2003, "Emotion detection in task-oriented dialogs corpus", *ICME*, Batimore.
- [9] L. Devillers, I. Vasilescu, C. Mathon, 2003, "Prosodic cues for perceptual emotion detection in task-oriented Human-Hum an corpus", *ICPhs*, Barcelone.
- [10] R. Fernandez, R. Picard, 2003, "Modeling Drivers' Speech Under Stress," *Speech Communication*.
- [11] C.M. Lee, S. Narayanan, R. Pieraccini, 2001, "Recognition of Negative Emotions from the Speech Signal", *ASRU*.
- [12] S. Narayanan, 2002, "Towards modeling user behavior in human-machine interactions: Effect of Errors and Emotions", *ISLE Workshop*, Edinburgh.
- [13] M. Lee et al., 2002, "Combining acoustic and language information for emotion recognition", *ICSLP*.
- [14] K. Sherer, 2003, "Vocal communication of emotion: A review and model for future research." *Speech Communication*.
- [15] K. Sherer, 1986, "Vocal affect expression: A review of research paradigms" *Psychological Bulletin*, 99, 143-165.
- [16] I. Murray JL. Arnott, 1993, "Toward the simulative of emotion in synthetic speech: a review of the literature on human vocal emotion", *Jasa*, 93/2, 1097-1108.