Tracking the Expression of Annoyance in Call Centers

Jon Irastorza and M. Inés Torres

Abstract Machine learning researchers have dealt with the identification of emotional cues from speech since it is research domain showing a large number of potential applications. Many acoustic parameters have been analyzed when searching for cues to identify emotional categories. Then classical classifiers and also outstanding computational approaches have been developed. Experiments have been carried out mainly over induced emotions, even if recently research is shifting to work over spontaneous emotions. In such a framework, it is worth mentioning that the expression of spontaneous emotions depends on cultural factors, on the particular individual and also on the specific situation. In this work, we were interested in the emotional shifts during conversation. In particular we were aimed to track the annoyance shifts appearing in phone conversations to complaint services. To this end we analyzed a set of audio files showing different ways to express annoyance. The call center operators found disappointment, impotence or anger as expression of annoyance. However, our experiments showed that variations of parameters derived from intensity combined with some spectral information and suprasegmental features are very robust for each speaker and annoyance rate. The work also discussed the annotation problem arising when dealing with human labelling of subjective events. In this work we proposed an extended rating scale in order to include annotators disagreements. Our frame classification results validated the chosen annotation procedure. Experimental results also showed that shifts in customer annoyance rates could be potentially tracked during phone calls. [1]

Index Terms: speech processing, emotion detection on Speech, annoyance tracking, machine learning, call center

---

Jon Irastorza
Speech Interactive Research Group, Universidad el País Vasco UPV/EHU (Spain), e-mail: jirastorza004@ehu.es

M. Inés Torres
Speech Interactive Research Group, Universidad el País Vasco UPV/EHU (Spain), e-mail: manes.torres@ehu.es

[1] This paper is a revised and extended version of a paper that was presented in [1].

1 Introduction

In recent year, the machine learning community has paid increasing attention to model the emotional status based on parameters derived from the analysis of the voice, the language, the face, the gestures or the ECG [2]. But data-driven approaches need corpora of human spontaneous behavior annotated with emotional labels [2] [3], which is a challenging requirement mainly due to the subjectivity of emotion perception by humans [2] [4]. As a consequence much research is being carried out over corpora that cover simulated or induced emotional behavior [4] [5]. It is worth mentioning that the selection of the situation where spontaneous emotions can be collected strongly depends on the goals of the research to be carried out. In particular the emotion identification from speech shows a wide range of potential applications and research objectives [6] [7]. Some examples are early detection of Alzheimer's disease [8], the detection of valency onsets in medical emergency calls [3] or in Stock Exchange Customer Service Centres [2]. It is clear that different sponteneous emotions arise in each of the previous situations.

When we deal with the recognition of emotions from speech signals we find a number of short-term features such as pitch, excitation signals, vocal tract features such as formants [9] [10], prosodic features [11] such as pitch loudness, speaking rate, rhythm, voice quality and articulation [3] [12], latency to speak, pauses [13] [14], features derived from energy [5] as well as feature combinations, etc.

Some surveys dealing with databases, classifiers, features and also with the set of categories to be identified in the analysis of emotional speech can also be found in the literature [10] [15] [16] [17]. Regarding methodology, classical classifiers such as the Bayesian or SVM have been proposed to analyze the feature distributions. The model of continuous affective dimensions is also an emerging challenge when dealing with continuous rating of emotion labelled during real interaction [18]. In this approach recurrent neural networks have been proposed to integrate contextual information and then predict emotion in continuous time to just deal with arousal and valence [19] [20].

However, once again further performance of any set parameters chosen or each identification method proposed depends of the research question to be addressed. In this work we addressed a research question proposed by a Spanish call-center providing customer assistance [13]. The goal of the company was to automatically detect annoyance rates during customer calls for further analysis, which resulted in a very challenge and novel goal. Their motivation was to verify if the policies that the company proposed to be implemented by operators when they have to deal with annoyed and angry customers really lead to shifts in the customer behavior. Thus an automatic procedure to detect those shifts would allow the company to evaluate their policies through the analysis of the recorded audios. Moreover their proposed to provide the operators with this information in real time, i.e. during the conversation. As a consequence our final work was aimed at tracking shifts in customer annoyance during conversations to complain services. To begin with, a short number of subjects showing very different ways to express their annoyance when complaining about a service were analyzed in this work. It is worth mentioninig that the call center op-

erators found disappointment, impotence or anger as expression of annoyance. We also discussed the annotation problem arising when dealing with human labelling of such subjective events. In this work we proposed an extended rating scale in order to include annotators disagreements. Then a certain amount of features were evaluated as potencial hints to track annoyance degrees through parametric and geometric classifiers. Intensity values and derivatives along with their corresponding suprasegmental values combined with some spectral analysis have demonstrated to be very robust in the experiments carried out over all the expressions of annoyance analyzed. Our results validated the proposed annotation procedure. The experimental results also showed that shifts in customer annoyance rates could be potentially tracked during phone conversations. In summary, the problem addressed in this work is the analysis of the intra-cognitive communication [21] [22] between customers and operators of a customer assistant service. In this framework the analysis carried out through the feature discussed can be seen as an artificial cognitive capability that actually measures a Human cognitive load.

2 Expression of Annoyance and Human Annotation

In this section, we describe the difficulties in get a consensus in both the human expression and the human perception of emotions. In particular we analyze the topics over a set of phone conversations aimed to complain about services provided by phone and internet companies.

2.1 Annotation Perception Difficulties

Due to the perceptual differences between people, the annotation of emotions is a very challenging problem to be addressed. Perceptual difficulties are caused by factors related to the culture [23] and the environment where we live as well as beliefs and experiences in our mental and maturity growth stage. In addition, people with psychiatric disorders like autism have more problems to understand and recognize emotion [24] [25]. These challenges are present primarily when we try to distinguish emotions with similar valence and arousal shown in figure 1. Nevertheless, we find more problems in distinguishing degrees within an emotional pure state. For instance, if we take annoyance as emotional pure state, low angry, medium angry and high angry could be three degrees to distinguish. In section 2.3.3 we can observe real difficulties annotating degrees of annoyance.

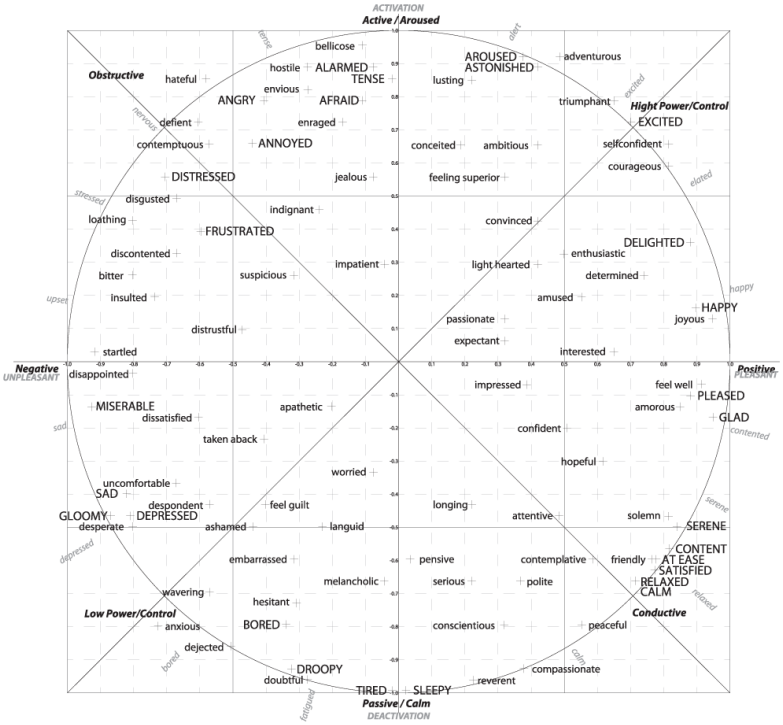


Fig. 1 Two-dimensional circumplex space model. Figure taken from Paltoglou et al. [26].

2.2 First Annotation Procedure

2.2.1 Expressing Annoyance During Phone Calls

Often, we find irate customers who call Telecommunications services for a variety of reasons because of slower broadband, landline fault or bill problem, among other things. The Spanish call center company involved in this work offers customer assistance for several phone, tv and internet service providers. The customer complaint services of these companies receive a certain number of phone-calls from angry or annoyed customers. But the way of expressing annoyance is not the same for all the customers. Some of them are furious and shout; others speak quickly with frequent and very short micropauses but do not shout [13], others seems to be more fed-up than angry; others feel impotent after a number of service failures and calls to the customer service.

Table 1 Expert annotations labels and duration of each audio analyzed

Expert annotation labels	Duration	Customer Speech Duration
Disappointed	00:00:42	00:00:31
Angry 1	00:00:42	00:00:10
Angry 2	00:00:35	00:00:24
Extremely Angry	00:16:20	00:07:50
Fed-up	00:01:08	00:00:20
Impotent	00:01:02	00:00:24
Annoyed in disagreement	00:01:35	00:00:29

2.2.2 Call Center Annotation Procedure

Due to the high number of recordings and issues regarding privacy, among other things, many audios were analyzed by experienced Call Center operators. These latter drew up a dataset and seven conversations with angry situations were selected and annotated as very annoyed customers. The customers show discomfort and disagreement because of service failures. In a second step, each conversation was named with the specific level of annoyance showed by the customer. Thus call center operators qualified the seven subjects in conversations as follows: *Disappointed*, *Angry* (2 records), *Extremely angry*, *Fed-up*, *Impotent* and *Annoyed in disagreement*. All these feelings correspond to the different ways the customer in the study expressed their annoyance with the service provided. More specifically they correspond to the way the human operators perceived customer feelings. Table 1 shows the duration of the conversations.

2.3 Second Annotation Procedure

Call center operators annotated each full audio file with one label. However we wanted to annotate segments inside each audio file since our final goal is to detect shifts in emotional degree appearing during the conversation. Thus we carried out the following segment-level annotation process.

2.3.1 Annotation Tool

The second annotation procedure was carried out using the *Praat* [27] software tool, which is a package for the scientific analysis of speech in phonetics designed to record, analyze synthesize and manipulate speech signals. Figure 2 shows a screenshot of the annotation procedure using *Praat* where from the top down we can observe four tiers representing respectively the audio waveform, the spectrogram joined with acoustic features, the annoyance numbered scale and the annoyance worded scale.

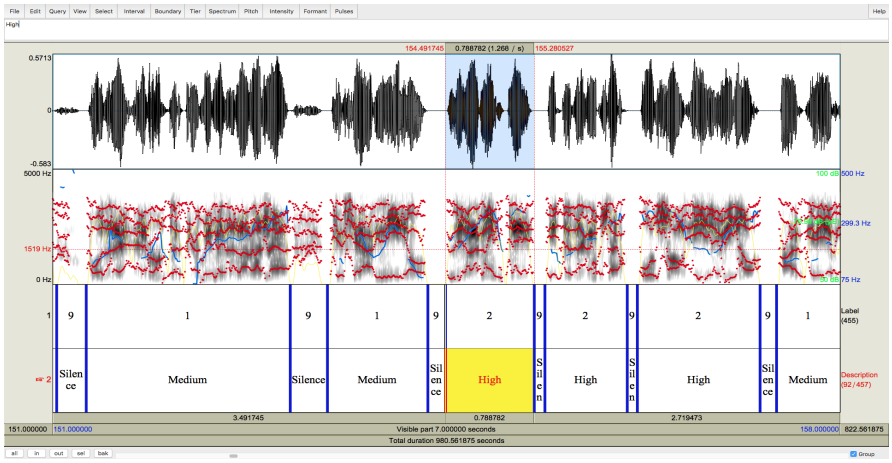


Fig. 2 Annotation example of 20 seconds interval of *Very Angry* audio using *Praat*. From the top down oscillogram tier, spectrogram tier, label tier and label-description tier are represented respectively.

2.3.2 Annotating Degrees of Annoyance

For this second annotation procedure, two Spanish male members of the research group [13] [28] annotated manually the segments. The task was done separately by each annotator in order to analyze and join annotations later. The age of each participant was 25 and 26 years respectively. Both Spanish annotators were graduate students.

First of all, audios were divided marking time steps into discrete segments such as customer speech -valid for these experiments-, operator speech, silence and overlapping. Then, annotators were asked to locate the changes of annoyance from the customer giving just one instruction: they had to identify up to three different levels of annoyance: zero for neutral or very low, one for medium and two for high degree. Table 2 shows the annotation labels chosen.

Table 2 Segment annotation levels

Label	Description
0	Neutral or Very Low
1	Medium
2	High
	Silence
	Operator or Overlapping segments

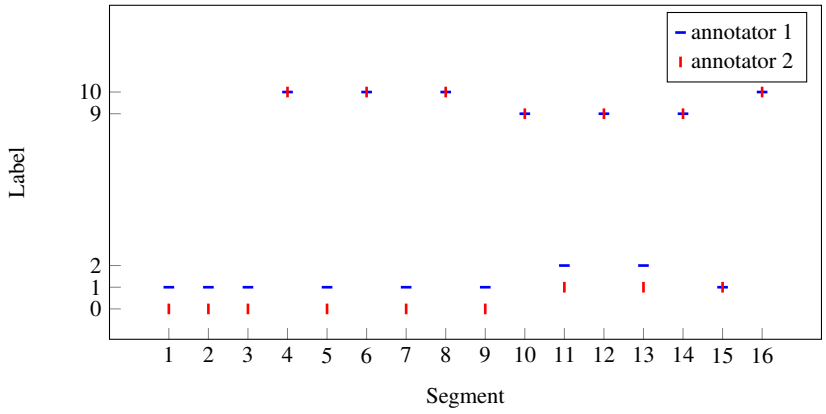
Table 3 Level scale defined

Ann. 1 label	Ann. 2 label	Final category
0	0	Very Low
0	1	Low
1	1	Medium
1	2	High
2	2	Very High

2.3.3 Annotation Difficulties

In the annotation procedure, a high level of disagreement between both annotators was obtained. Annotating and classifying emotions is a complex task, where usually a significant level of disagreement is present due to context, individual and cultural factors [4] that create an environment of subjectivity. Moreover, in this labeling process the annotators had to identify between different levels of anger, which is a further difficulty because of the different ways to expressed anger by each customer analyzed. The behavior of each annotator to annotate, resulted many disagreements. Second annotator, tended to perceive one lower degree than the first one. For this reason, we can observe many segments annotated as zero/one and one/two, while one/zero and two/one were rare. Figure 3 shows graphically disagreements between annotators. This issue between annotators appeared because the second-one paid more attention to the words and the first one paid more attention to the tone of voice. These reasons could explain why some number of zero/two were present in *Very Angry* audio file. However, the annotator agreement was high in the identification of the time steps where they perceived changes in the degree of expression. Then, just one of them was chosen to fix segments bounds.

Fig. 3 Disagreements between annotators in *Disappointed* audio



2.3.4 Defining Level Scale

At this point we needed to arrange disagreements defining an extended scale of annoyance expression. On this scale we have three agreed degrees corresponding to *very low*, *medium* and *very high*. The latter degree is only present in the audio entitled as *very angry*. Table 3 shows segment annotation label. Furthermore, we have two disagreed degrees corresponding to *low* and *high*, which correspond to *low-medium* and *medium-high* disagreements respectively. Table 4 shows the final number of segments and frames for each audio file and annoyance level. In this

Table 4 Number of segments and frames for each audio file

	Activation level category											
	v_low		*low*		*medium*		*high*		*v_high*		*All*	
	Segments	*Frames*	*Segments*	*Frames*	*Segments*	*Frames*	*Segments*	*Frames*	*Segments*	*Frames*	*Segments*	*Frames*
Disappointed	0	0	5	2095	1	456	1	3495	0	0	7	6046
Angry 1	0	0	0	0	5	1545	1	346	0	0	6	1891
Angry 2	1	412	2	1471	1	499	1	2445	0	0	5	4827
Very angry	0	0	27	9421	40	15922	97	43026	45	19221	209	87590
Fed-up	0	0	3	783	11	3018	1	334	0	0	15	4135
Impotent	4	2600	4	1432	2	732	0	0	0	0	9	4764
All	5	3012	41	15202	59	22172	101	49646	45	19221	251	109253

Table, the audio *annoyed in disagreement* does not appear because of not changes on the customer rate of annoyance during the conversation.

3 Features

In this section, we describe the primary features used as well as the procedure and the features package selected. Primary features extraction, as well as the annotation procedure described in Section 2.3, was carried out using the *Praat* [27] software tool. In this case, scripts were written and used to extract the features desired. Furthermore, we wrote scripts to calculate normalized and derivatives values.

3.1 Primary Features

For this work, we used four primary features like Pitch, Intensity, Mel-Frequency Cepstral Coefficients (MFCC) and Linear Prediction Coefficients (LPC). The Pitch represents the fundamental frequency (f_0) of speech signal. The intensity is defined as the energy per unit area transmitted by the acoustic wave. The MFCC is based on frequency domain using the Mel scale and is a representation of the real cepstral of a windowed short-time signal derived from the Fast Fourier Transform (FFT) of that signal. The LPC coefficients are calculated by LPC estimation which describes the inverse transfer function of the human vocal tract.

3.2 Procedure

In these experiments, we chose a short-term and a long-term analysis of speech signal using 20 ms overlapping window and 330 ms window [28] respectively. This

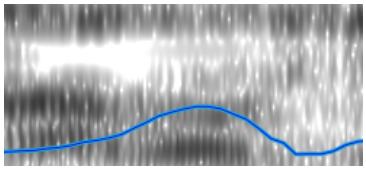


Fig. 4 Pitch feature represented in *Praat*

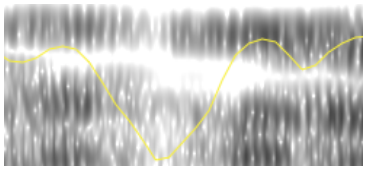


Fig. 5 Intensity feature represented in *Praat*

period roughly corresponds to the word utterance level for Spanish tongue. In our work it includes 66 overlapped frames each five miliseconds. Overall, we extracted 350 values per frame such as primary, normalized, derivative and suprasegmental parameters.

On the one hand, as primary local features, we calculated 35 parameters: Pitch, Intensity, 5 Formants, 12 Mel-Frequency Cepstral Coefficients (MFCC) and 16 Linear Prediction Coefficients (LPC). Then we also calculated 35 normalized values using the mean from all the speaker frames and 70 derivative values -first and second derivatives- by comparing the local value with the following frame.

On the other hand, as suprasegmental features, we calculated 210 parameters: the smoothed value, i.e. mean value over the suprasegmental analysis window, smoothed first and second derivatives, the standard deviation and the standard deviation of the first and second derivatives.

4 Experimental Evaluation

The experiments carried-out had a two-fold objective: one was to analyze the validity of the assumptions made in the annotation procedure, i.e. the defined set of classes, the other was to select a set of discriminative parameters for further frame classification. To this end two sets of experiments were carried out. In the first one we got speaker dependent results, whereas in the second one all the audios were included in global experiments.

4.1 Experimental Framework

Several tools are used to analyse and experience such as Accord Framework, CUDA or Scikit-Learn. This latter was used in these experiments. More specifically, we used a parametric Nave Bayes Classifier (NB) and the Support Vector Machine (SVM), which is a no parametric learning model. We also used k-Nearest Neighbors (k-NN) to compare with SVM due to the results of the latter.

4.1.1 Classifiers

Naïve Bayes Classifier assumes that features are conditionally independent but they are distributed according with some parametric distribution whose parameters have to be estimated during the learning procedure, typically using a Maximum Likelihood Estimate (MLE). Bayes theorem states the following relationship:

$$P(y|x_1,...,x_n) = \frac{P(y)P(x_1,...,x_n|y)}{P(x_1,...,x_n},$$

where **y** represents a class variable and $\mathbf{x_1}$ through $\mathbf{x_n}$ a dependent feature vector. In our use case, we have chosen to use the Gaussian Naive Bayes algorithm for classification. The likelihood of the features is assumed to be Gaussian:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}).$$

On the other hand geometric classifiers do not assume any distribution of data and focus on searching boundaries. In particular Support Vector Machine (SVM) looks for hyperplanes separating data in different classes in high dimensional spaces, whereas K-NN classification is based on distance between class prototypes. The disadvantage of the SVM classifier with respect both NB and k-NN classifiers is the time complexity. An SVM is a quadratic programming problem that can scale between:

$$O(n_{features} \times n_{samples}^2) \text{ and } O(n_{features} \times n_{samples}^3).$$

For this reason, the compute and storage requirements increase depending on the dataset.

4.1.2 Experiments

Table 4 show us the number of segments and set of frames used. Categorizing and segmenting speech segments in different degrees of annoyance are difficult and subjective tasks, making the frame classification process which is our goal in this work even more complex.

Then two series of classification experiments were carried out using two different sets of features. To this end we shuffled the set of frames of each audio file and then split this set into a training and a test set that included 70% and 30% of frames, respectively. In both series of experiments we used the frame classification accuracy as evaluation metric. Then we also calculated precision, recall and F-measure values for each category defined in Section 2.3. The metrics have been selected to evaluate the quality of the frame classification. In both Section 4.2.1 and Section 4.2.2 we used accurary and F-measure as most representative evaluation metrics of each experiments. F-measure is defined as follows:

Table 5 Frame classification accuracy for each set of features and customer audio file, using SVM and NB

		LPC	*MFCC*	*Formants*	*Intensity*	*dIntensity*	*Pitch*	*d_Pitch*	*Primary*
SVM	*Disappointed*	0.61	0.60	0.58	0.60	**0.84**	0.60	0.61	0.57
	Angry 1	0.81	0.81	0.80	0.82	**0.87**	0.81	0.82	0.80
	Angry 2	0.60	0.51	0.50	0.50	**0.80**	0.56	0.54	0.50
	Very angry	0.47	0.47	0.47	0.47	**0.56**	0.47	0.48	0.47
	Fed-up	0.74	0.72	0.72	0.71	**0.83**	0.73	0.75	0.74
	Impotent	0.58	0.55	0.54	0.57	**0.76**	0.55	0.59	0.54
NB	*Disappointed*	0.56	**0.62**	0.59	0.59	0.60	0.57	0.59	0.55
	Angry 1	0.65	**0.82**	0.78	0.79	0.81	0.80	0.85	0.72
	Angry 2	0.35	**0.56**	0.54	0.51	0.49	0.52	0.36	0.50
	Very angry	0.44	**0.48**	0.47	0.46	0.45	0.47	0.37	0.40
	Fed-up	0.63	**0.75**	0.71	0.72	0.65	0.73	0.37	0.69
	Impotent	0.51	**0.56**	0.55	0.53	0.55	0.54	0.38	0.52

$$F = 2.\frac{recall \times precision}{recall + precision}.$$

4.2 Speaker Dependent Experiments

4.2.1 First Series of Experiments

The goal of this series of experiments was to analyze the performance from the previous cited three classifiers using the selected sets of features and classifying annoyance degrees for each speaker.

Firstly, SVM and NB models were used to classify frames in order to evaluate the following set of features: LP and MFC coefficients, Formants, Pitch, Intensity and two more sets adding the normalized value, the first and second derivatives and the suprasegmental values to Pitch and Intensity. The latter two sets resulting in a total of 10 values and are referenced in Figures and Tables as dPitch and dIntensity. We also tested the set Primary of the 35 primary local parameters. In Table 5 we can see the frame classification accuracy for each set of features, speaker and classifier.

We can see the best results achieved when the SVM classifier and the set dIntensity were used. The highest and the lowest accuracies were obtained in *Angry 1* and *Very Angry* audio files with 0.87 and 0.56 respectively. Suprasegmental features seem to be fairly reliable when annoyance degrees are classified. The other sets of features showed worse performance except for the *Angry 2* audio file whose results we included here can be considered only a little worse. Furthermore, LPC show slightly better performance than MFCC and dPitch unimproved Pitch performance.

Then again, the NB classifier show worse performance and the best classification accuracies for all audios were achieved by MFCC, suggesting that these spectral features matches better the parametric distribution assumption of NB classifier.

Finally, we can se in Table 5 highest accuracy results for customer entitled as *Angry 1* and lowest accuracy results for customer entitled as *Very angry*.

4.2.2 Second Series of Experiments

Secondly, SVM, NB and k-NN classifiers were evaluated using the following set of features: LP and MFC coefficients, dIntensity and dIntensity combined with LPC and MFCC. We compared both SVM and NB and both SVM and k-NN classifiers. The results obtained are represented in Figures 6 and 7 where bar graphs show the frame identification accuracies per each of features selected for each audio file. Figure 6 shows the frame classification accuracy for each audio and annoyance rate, when both SVM and NB models were used to classify the selected sets of features. It confirms that classification performance achieve better results classifying with SVM. The highest accuracy results were obtained combining spectral information from LP coefficients and suprasegmental information -LPC plus dIntensity features- for all audio files analyzed. The worst accuracy results were obtained with *Very Angry* achieving 0.64. In the rest of the audios we obtained an accuracy higher than 0.85 for all the speakers. The overall accuracy of this classifier with dIntensity features was 0.68 versus 0.45 of a majority guess baseline. The annoyance degree classification in this model can be measured to an accuracy better than 20 per cent for the *Angry 1* audio file and 53 per cent for the *Angry 2* one.

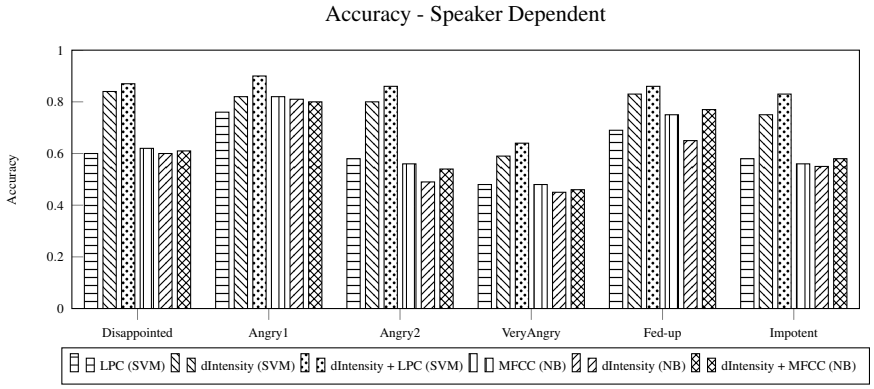


Fig. 6 Comparison of SVM and NB frame classification approaches. The three bar graphs on the left side of each audio file (blue, red and brown) correspond to results obtained by SVM classifier whereas the ones of the right side (grey, purple and green) corresponds to the results obtained by NB.

In order to extend our study, we decided to compare SVM results with a distance-based classifier. Thus, we chose K-Nearest Neighbors Classifier wich is a nonparametric classification algorithm that find a predefined number of k training samples closest in distance to the new point, and assigns their predominant label. Figure 7 shows highest accuracies values with k-NN for k=1 when MFCC features were added to the dIntensity ones. K-NN classifier obtained an accuracy of 0.91 for *very angry* audio file and an accuracy around 0.98 for al the others customers. It should be noted that we are trying to classify segments at frame-level. For this reason, it is easy to find the closest prototype and we found the SVM classifier more robust and reliable.

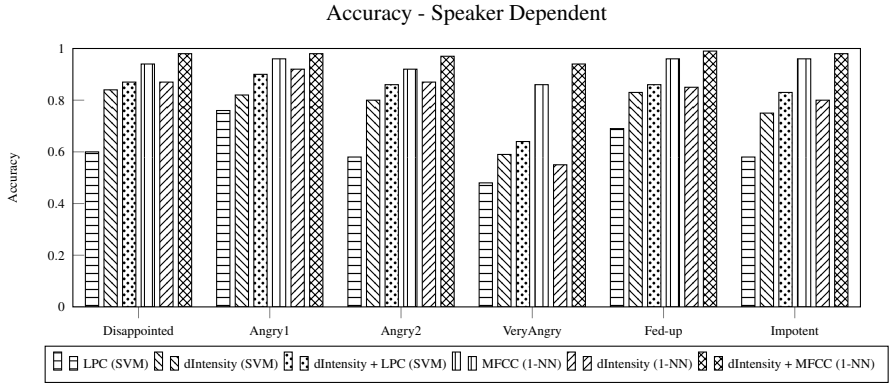


Fig. 7 Comparison of SVM and k-NN (k=1) frame classification approaches. The three bar graphs on the left side of each audio file (blue, red and brown) correspond to results obtained by SVM classifier whereas the ones of the right side (grey, purple and green) corresponds to the results obtained by k-NN.

Figure 8 to Figure 12 show the frame classification F_measure obtained by each annoyance rate for each audio file, when SVM model was used to classify the selected sets of features. Bar graphs in the Figures show the frame identification F_measure for each class, based on each of the sets of features selected for these experiments. They confirm that the set of dIntensity along with the LPC values led to the best frame classification results for every activation rate. Figure 9 and 10 also show poor classification rates for *Low* and *Medium* activation rates in *very angry* audio file, which explains the lower overall frame classification accuracy obtained in this audio file (see Table 5 and Figure 6). Finally, we can see the importance of the suprasegmental analysis included in the intensity analysis when shifts have to be detected. However the combination with spectral features led to even higher identification rates both in terms of Accuracy and F-measure.

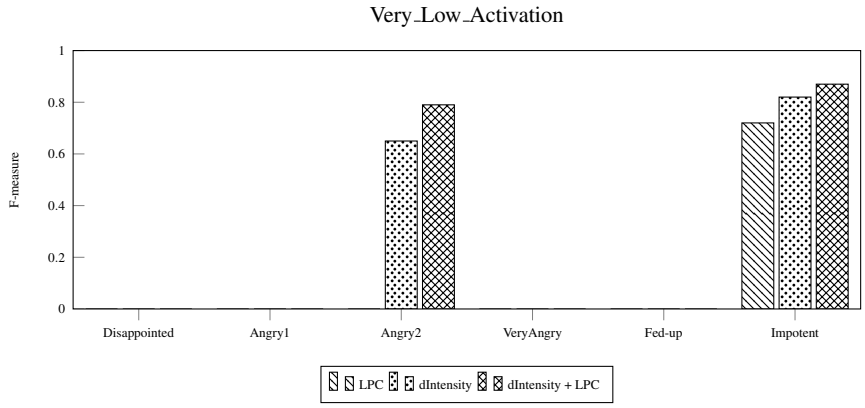


Fig. 8 Comparison of SVM frame classification approach for *Very Low* activation value.

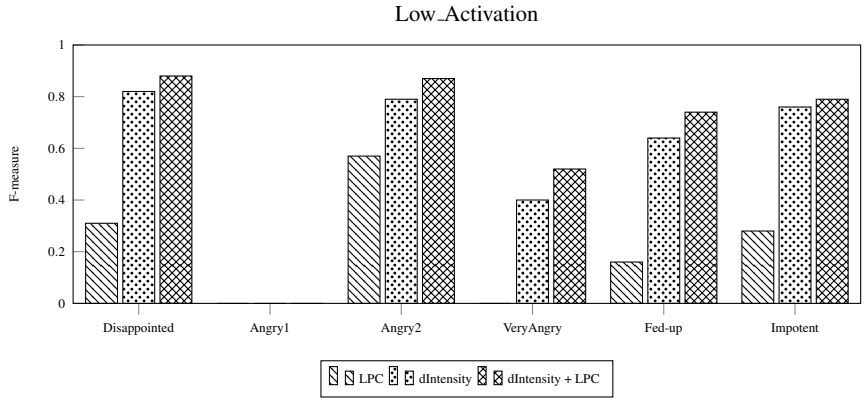


Fig. 9 Comparison of SVM frame classification approach for *Low* activation value.

4.3 Global Experiments Results

Global experiments was aimed at analyzing the behavior of the best sets of features in previous experiments (see Figure 6) when all the frames in audio files were shuffled to be considered in a global experiment. To this end we split the whole set of frames into ten folders to carry out a 10-Cross Validation evaluation procedure. The whole sample set included 109253 frames. Silence frames were not considered in this task.

Figure 13 shows lower classification rates than the previous speaker dependent experiments in Figure 6. However, similar behavior of the set of features and classifiers is observed in this case. The highest frame classification accuracy, 0.95, was obtained when K-NN classifier was used for *dIntensity + MFCC* feature vector. K-NN classifier continues giving excellent results due to the reasons explained in

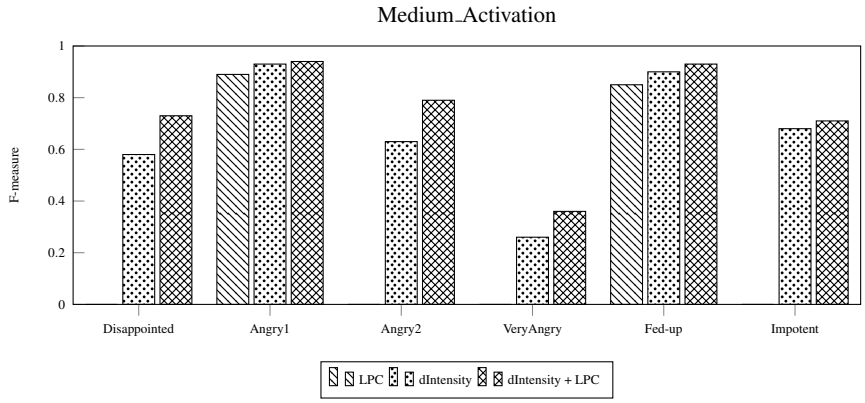


Fig. 10 Comparison of SVM frame classification approach for *Medium* activation value.

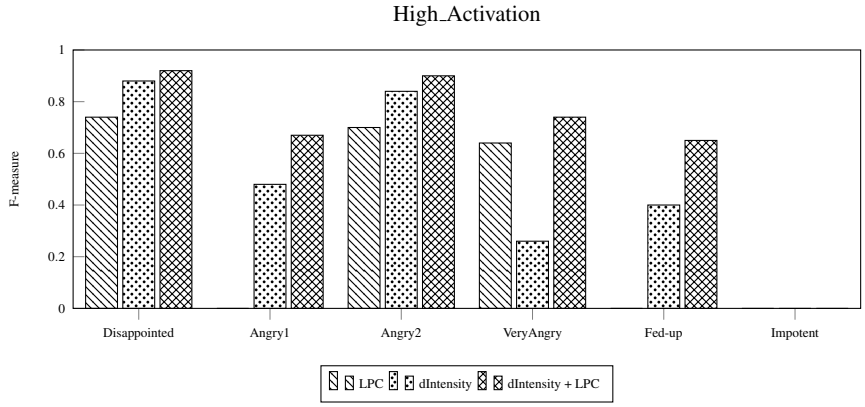


Fig. 11 Comparison of SVM frame classification approach for *High* activation value.

Section 4.2.2. Anyway, in Figure 14 we can observe F-measure scores around 0.7 and 0.8 using SVM and NB classifiers respectively for *Very Low*.

Figure 14 shows the F-measure obtained by both SVM and K-NN classifiers for each category representing the global anger degree defined in Section 2 through a cross-validation evaluation procedure of the sets of features in Figure 6.

5 Conclusions

In this work we addressed the detection and the of the spontaneous expression of annoyance during real conversations between customers and operators of phone customer services, which is a really novel and challenging goal. Our final work has been to identify and tracking shifts in customer annoyance during the phone-calls,

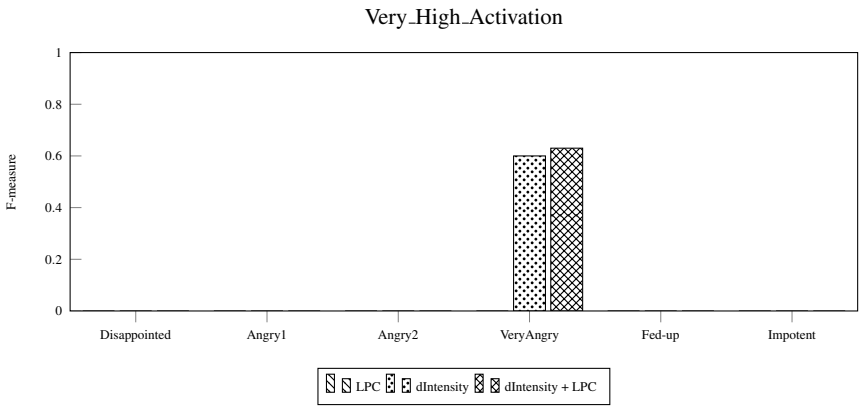


Fig. 12 Comparison of SVM frame classification approach for *Very High* activation value.

in Spanish. To this end we have analyzed a set of audio files showing different ways to express the annoyance or anger of the customer. Some of them were furious and shouted; others spoke quick with frequent and very short micropauses but did not shout [13], others seemed to be more fed-up than angry; others felt impotent after a number of service failures and calls to the customer service. A number of seven conversations were analyzed in the work. In them the call center operators found disappointment, impotence or anger as expression of annoyance in these audio files. The work has also discussed the annotation problem arising when dealing with human labelling of subjective events. In this work an extended rating scale has been proposed in order to include annotators disagreements. Then a certain amount of features were evaluated as potencial hints to track annoyance degrees through parametric and geometric classifiers. Local features including acoustic and prosodic parameters, their normalized values, derivatives and a set of suprasegmental parameters have been extracted. Intensity and intensity-based suprasegmental features has shown to be very robust to identify class boundaries in every audio file analyzed. A combination of intensity-based suprasegmental features with LPC coefficients led to the best frame classification accuracies for all the expressions of annoyance analyzed for SVM classifier, whereas the combination with MFCC coefficients got the best results when K-NN classifier was used.

The obtained frame classification results validated the chosen annotation procedure. However it should be extended to a higher number of both conversations and annotators so that the procedure could be adjusted and improved. Experimental results also showed that shifts in customer annoyance rates could be potentially tracked during phone calls.

One of the main goals of this study was design and implement a tool capable of monitoring the different customer annoyance degrees. The annotation process is not an easy task because of the different issues mentioned in Section 2.1. These factors appear when annotators begin to annotate data as can be proven by Section 2.3.3. Thus, the subjectivity of this process is directly reflected in recognition per-

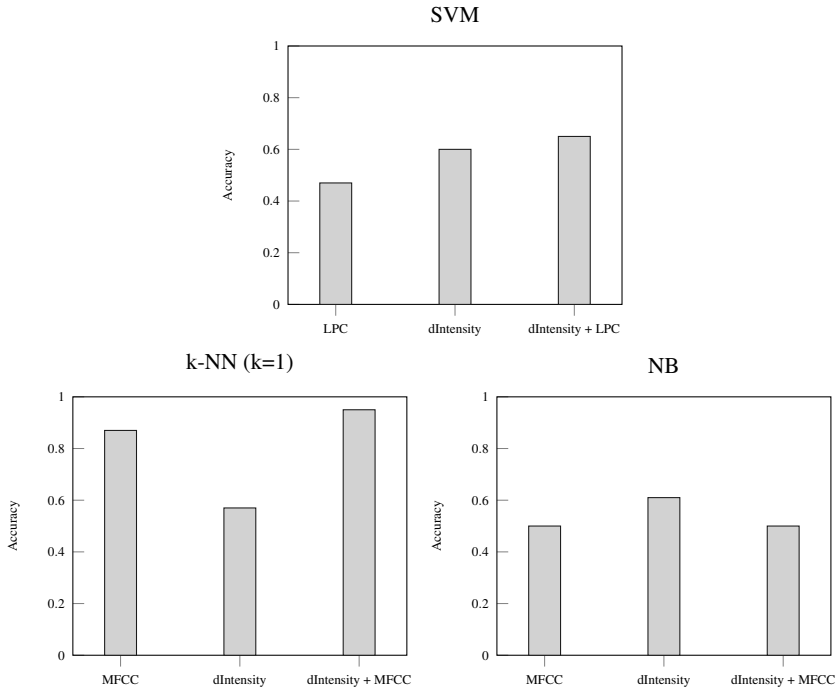


Fig. 13 Frame identification accuracy for each category and feature vector when SVM, k-NN and NB classifiers were used in a global cross validation experiment where all frames extracted from all audio files were included

formance because the input subjectivity will generate output subjectivity. We can try to solve this problem in some environment cases or cases where the subject analyzed is known thanks to the analyst cognitive capacity. For this reason, at this early stage, we can consider the synergy between the Decision Support System (DSS) and the call center in order to gain objectivity and later adjust the DSS.

The problem addressed in this work is a good example to show synergies between humans and artificial cognitive systems, between engineering and cognitive sciences [22]. In particular the behavior of the classifiers to identify different activation levels at frame level and their ability to identify shifts in annoyance level can be interpreted as a measure of the human cognitive load when dealing with the same problem.

6 Acknowledgements

This work has been partially founded by the Spanish Mineco under grant TIN2014-54288-C4-4-R and by the H2020 EU under Empathic RIA action number 769872.

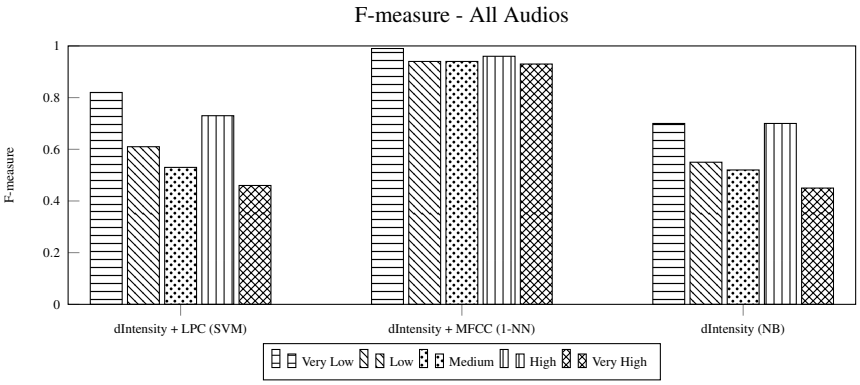


Fig. 14 F-measure values computed for each category in this second series of experiments by SVM, k-NN and NB classifiers through the feature sets in Figure 6 and a cross validation procedure

References

1. J. Irastorza and M. I. Torres, "Analyzing the expression of annoyance during phone calls to complaint services," in *Cognitive Infocommunications (CogInfoCom), 2016 7th IEEE International Conference on*. IEEE, 2016, pp. 000 103–000 106.
2. L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407 – 422, 2005, emotion and Brain.
3. L. Vidrascu and L. Devillers, "Detection of Real-Life Emotions in Call Centers," in *Proceedings of INTERSPEECH'05: the 6th Annual Conference of the International Speech Communication Association*. Lisbon, Portugal: ISCA, 2005, pp. 1841–1844.
4. S. E. Eskimez, K. Imade, N. Yang, M. Sturge-Apple, Z. Duan, and W. Heinzelman, "Emotion classification: how does an automated system compare to naive human coders?" in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, March 2016, pp. 2274–2278.
5. J. C. Kim and M. A. Clements, "Multimodal affect classification at various temporal lengths," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 371–384, Oct 2015.
6. M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '14. New York, NY, USA: ACM, 2014, pp. 3–10.
7. C. Clavel and Z. Callejas, "Sentiment analysis: From opinion mining to human-agent interaction," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 74–93, Jan 2016.
8. J. J. G. Meilán, F. Martínez-Sácnhez, J. Carro, D. E. López, L. Millian-Morell, and J. M. Arana, "Speech in alzheimer?s disease: Can temporal and acoustic parameters discriminate dementia?" *Dementia and Geriatric Cognitive Disorders*, vol. 37, no. 5-6, pp. 327–334, 2014.
9. K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using fourier parameters," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 69–75, Jan 2015.
10. D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162 – 1181, 2006.
11. B. M. Ben-David, N. Multani, V. Shakuf, F. Rudzicz, and P. H. H. M. van Lieshout, "Prosody and semantics are separate but not separable channels in the perception of emotional speech: Test for rating of emotions in speech," *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 1, pp. 72–89, 2016.

12. J. M. Girard and J. F. Cohn, "Automated audiovisual depression analysis," *Current Opinion in Psychology*, vol. 4, pp. 75 – 79, 2016.
13. R. Justo, O. Horno, M. Serras, and M. I. Torres, "Tracking emotional hints in spoken interaction," in *Proc. of VIII Jornadas en Tecnología del Habla and IV Iberian SLTech Workshop (IberSpeech 2014)*, 2014, pp. 216–226.
14. A. Esposito, A. M. Esposito, L. Likforman-Sulem, M. N. Maldonato, and A. Vinciarelli, *Recent Advances in Nonlinear Speech Processing*. Cham: Springer International Publishing, 2016, ch. On the Significance of Speech Pauses in Depressive Disorders: Results on Read and Spontaneous Narratives, pp. 73–82.
15. M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572 – 587, 2011.
16. B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 910, pp. 1062 – 1087, 2011.
17. C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155–177, 2015.
18. A. Mencattini, E. Martinelli, F. Ringeval, B. Schuller, and C. D. Natlae, "Continuous estimation of emotions in speech by dynamic cooperative speaker models," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, 2016.
19. M. Wollmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," 9 2008, pp. 597–600.
20. F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters*, vol. 66, pp. 22 – 30, 2015.
21. P. Baranyi and A. Csapó, "Definition and synergies of cognitive infocommunications," *Acta Polytechnica Hungarica*, vol. 9, no. 1, pp. 67–83, 2012.
22. P. Baranyi, A. Csapó, and G. Sallai, *Cognitive Infocommunications (CogInfoCom)*. Springer International, 2015.
23. M. Koeda, P. Belin, T. Hama, T. Masuda, M. Matsuura, and Y. Okubo, "Cross-cultural differences in the processing of non-verbal affective vocalizations by japanese and canadian listeners," 2013.
24. K. M. Rump, J. L. Giovannelli, N. J. Minshew, and M. S. Strauss, "The development of emotion recognition in individuals with autism," *Child development*, vol. 80, no. 5, pp. 1434–1447, 2009.
25. C. Ashwin, E. Chapman, L. Colle, and S. Baron-Cohen, "Impaired recognition of negative basic emotions in autism: A test of the amygdala theory," *Social neuroscience*, vol. 1, no. 3-4, pp. 349–363, 2006.
26. G. Paltoglou and M. Thelwall, "Seeing stars of valence and arousal in blog posts," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 116–123, 2013.
27. P. Boersma and D. Weenink, "Praat: doing phonetics by computer," University of Amsterdam, Software tool, 2016, version 6. 0.15. [Online]. Available: http://www.fon.hum.uva.nl/praat/
28. M. Iturriza, "Identificacin de activacin emocional adaptada a cada locutor," *Graduation thesis. Universidad del País Vasco*, 2015.