MINING CALL CENTER CONVERSATIONS EXHIBITING SIMILAR AFFECTIVE STATES

Rupayan Chakraborty, Meghna Pandharipande, and Sunil Kumar Kopparapu
TCS Innovation Labs-Mumbai
Yantra Park, Thane West-400601, India
{rupayan.chakraborty, meghna.pandharipande, sunilkumar.kopparapu}@tcs.com

Abstract

Automatic detection and identifying emotions in large call center calls are essential to spot conversations that require further action. Most often statistical models generated using annotated emotional speech are used to design an emotion detection system. But annotation requires substantial amount of human intervention and cost; and may not be available for call center calls because of the infrastructure issues. Therefore detection systems use models that are generated form the readily available annotated emotional (clean) speech datasets and produce erroneous output due to mismatch in training-testing datasets. Here we propose a framework to automatically identify the similar affective spoken utterances in large number of call center calls by using the emotion models that are trained with the freely available acted emotional speech. Further, to reliably detect the emotional content, we incorporate the available knowledge associated with the call (time lapse of the utterances in a call, the contextual information derived from the linguistic contents, and speaker information). For each audio utterance, the emotion recognition system generates similarity measures (likelihood scores) in *arousal* and *valence* dimension using pretrained emotional models, and further they are combined with the scores from the contextual knowledge-based systems, which are used to reliably detect the similar affective contents in large number of calls. Experiments demonstrate that there is a significant improvement in detection accuracy when the knowledge-based framework is used.

Index Terms:Affective content analysis; mining call center audio; spontaneous emotional speech; knowledge-based systems; similar affective states

1 Introduction

Affective content [1] analysis of audio calls is important in recent days with the increasing number of call centers (Pang and Lee, 2008), (Liu, 2012), (Kopparapu, 2015). Perhaps, audio is the best possible modality that can be used to effectively analyze the call center conversations between customer and agent. However, manual analysis of such calls is cumbersome and may not be feasible because large number of recordings take place on daily basis. Therefore only a small fraction of such conversations are carefully heard by the human supervisors and addressed, thus resulting many of those unattended.

The difficulty of identifying the affective regions (or emotionally rich) manually in large number of calls is illustrated in Figure 1. The call duration is plotted on the x-axis, while different calls are shown along the y-axis. As represented in Figure 1, the calls are of different durations, and the gray color represents the actual length of the calls. The black color within the call shows the location of a specific affective state (highly correlated to the problematic regions in the calls). It is clear that the locations of such problematic regions are arbitrary, and the durations are of variable length. In spite of such challenges, automatic emotion analysis

[1] Affective content and Emotion will be used interchangeably in this paper

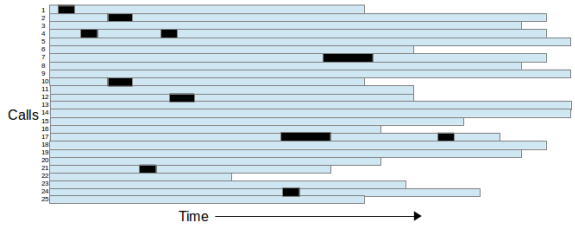

Figure 1: Call center calls (dark portions indicate problems)

of call center conversations has attracted the attention of researchers (for example (Petrushin, 1999), (Vidrascu and Devillers, 2007), (Gupta and Rajput, 2007), (Mishra and Dimitriadis, 2013), (Kopparapu, 2015), (Pappas et al., 2015)).

Affective content analysis is a technique that extracts emotions from spoken utterance [2], and thus useful to find the similar emotional utterances in call center calls. In general, affective contents are represented categorically in terms of the different emotion classes (e.g (Petrushin, 1999; Vidrascu and Devillers, 2007; Gupta and Rajput, 2007; Pappas et al., 2015)). Mostly in call center calls, four emotions (namely, *anger*, *happy*, *neutral*, *sad*) in the categorical space are addressed. Although in (Nicolaou et al., 2011), authors proposed to capture the time varying emotional information in the dimensional space using audio-visual cues. And in (Mishra and Dimitriadis, 2013), authors proposed an incremental emotion recognition system that updates the recognized emotion with each recognized word in the conversation. They make use of three features from two modalities (i.e. cepstral and intonation form audio and textual features from text), which are obtained at the word level to estimate the emotion with better accuracies. It has been observed that combining linguistic information with the acoustic features improves the performance of the system. As an example, in (Lee and Narayanan, 2005), authors proposed a combination of three information (i.e. acoustic, lexical, and discourse) for emotion recognition in spoken dialogue system and found improvements in recognition performance. Similarly in (Planet and Sanz, 2011), authors described an approach to improve emotion recognition in spontaneous children's speech by combining acoustic and linguistic features.

In this paper, we propose a novel framework that automatically extracts the affective content of the call center spoken utterances in *arousal* and *valence* dimensions. In addition, context-based knowledge (e.g. time lapse of the utterances in the call, events and affective context derived from linguistic content, and speaker information) associated with the calls are intelligently used to reliably detect the affective content in speech. Unlike (Mishra and Dimitriadis, 2013), we do not fully rely on the use of word recognition to determine the emotion. This makes our system feasible even for resource deficient languages that do not boast of a good automatic speech recognition (ASR) engine. In addition to the linguistic information like in (Lee and Narayanan, 2005), we also incorporate more knowledge like the time lapse of the utterance in calls, contextual information derived from linguistic content, speaker information etc. For each spoken utterance, the affective content extractor generates probability scores in *arousal* and *valence* dimensions, which are then probabilistically combined to label it with any of the predefined affective classes. The framework is motivated by the observation that there is significant disagreement amongst human annotators when they annotate call center speech; the disagreement largely reduces when they are provided with additional knowledge related to the conversation. Unlike (Mishra and Dimitriadis, 2013), the proposed system extracts affective information separately in dimensional space, thus reduces the classification complexity. Moreover in our proposed framework, emotions are extracted at discrete levels of affective classes (i.e. positive, neutral, negative in arousal and valence dimensions), instead of using affective information in continuous scale like in (Nicolaou et al., 2011), thus reducing the complexities related to the difficulties in annotation at continuous level of affective states, resulting less number classes in each dimension. In addition, detection of similar emotional content in large number of audio calls are performed by using the emotion models trained with the freely available acted emotional speech. Therefore the system is able to work even in a scenario if somebody does not have an annotated call center calls. Extensive experimentations on the acted dataset contaminated with 4

[2]The word utterance, turn, and spoken terms will be used interchangeably from here on

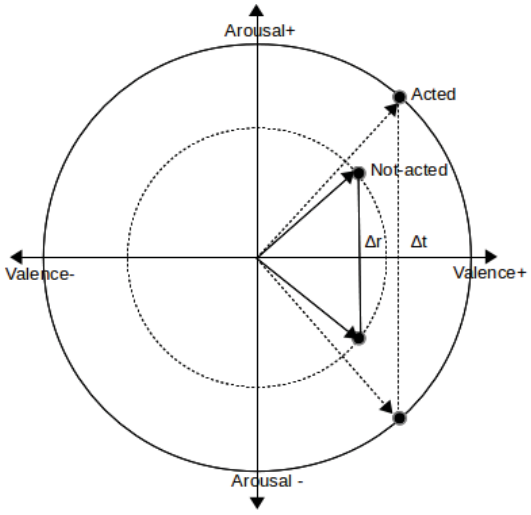

Figure 2: Error in emotion estimation for spontaneous call center calls

different types of noise (babble, F-16, machine-gun and volvo) from Noisex-92 dataset at different SNR levels has also been carried out. The main contributions of the paper are: (i) affective content extraction from large call center calls and also finding similar segments (ii) incorporation of the knowledge for reliable detection.

The rest of the paper is organized as follows. Section 2 presents the motivation of this work. In Section 3, we propose the framework for affective content extraction, using knowledge for reliable affective state identification and finding similar affective states. Section 4 describes the dataset, experiments, and results. We conclude in Section 5.

2 Motivation and challenges

Voice-based call center handle and record thousands of calls on daily basis. It is a difficult task for someone to identify manually the emotional segments from these large number of calls; and subsequently decide which of these recorded calls are to be selected for taking necessary actions regarding the issues related to customers dissatisfaction. In general, the calls are of variable duration, where time length varying from few seconds to few minutes depending on several factors, like the type of the call center, the type of the problems that the customers are facing, the prior affective state of the customer, the way agents handle the problem and behave etc. In such scenario, super agent (or supervisors) manually select at random few calls from a large number of calls, then listen and check if there are some abnormalities. In this way, very few calls are normally analyzed and addressed, thus resulting many calls unattended, ends up with a increasing dissatisfaction among the customers. Automatic detection of the equivalent emotional segments in large set of calls are useful to deal with such situation. Like in other pattern classification problems, most often the statistical models are generated using the annotated data and then used to detect the affective states in the audio calls.

It motivates us to think about a system that learns the pattern of different affective states from freely available acted emotional speech, and detect similar affective parts in large number of call center calls. Since, the classifiers trained and tested in different environments, are expected to give erroneous output. The difference of similar emotions in two different environments is elaborated in Figure 2 that represents the two affective dimensions of emotion, namely, *arousal* (also referred as activation) and *valence*. A point in this 2D space can be looked upon as a vector and is representative of an emotion. The acted speech in the training dataset exhibits higher degree of intensity, both in *arousal* and *valence* dimensions resulting in a larger radii emotion vector compared to the spontaneous speech. On the other hand, call center spontaneous speech has lesser intensity than the acted speech. For this reason, it is easy to mis-recognize one emotion for another in call center speech. Subsequently, if the first quadrant (Figure 2) represents emotion E_1 and the fourth quadrant represents emotion E_2, then the mis-recognition error is small (Δr) for call center speech but requires higher degree of error in judgment (Δt) to mis-recognize emotion E_1 as emotion E_2 and vice-versa for acted speech.

To handle such mismatch in train-test environments, the proposed system intelligently makes use of several knowledge for the reliable extraction of the affective content in call canter calls. For an audio segment in a call, the idea is to generate a probability matrix, whose elements are a joint probability estimate in the affective dimension of *arousal* and *valence* . Then all the knowledge-based information are used to modify the elements of the matrix. Simi-

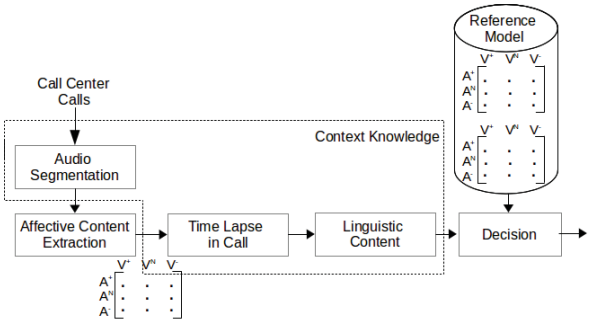

Figure 3: Call center call analysis framework

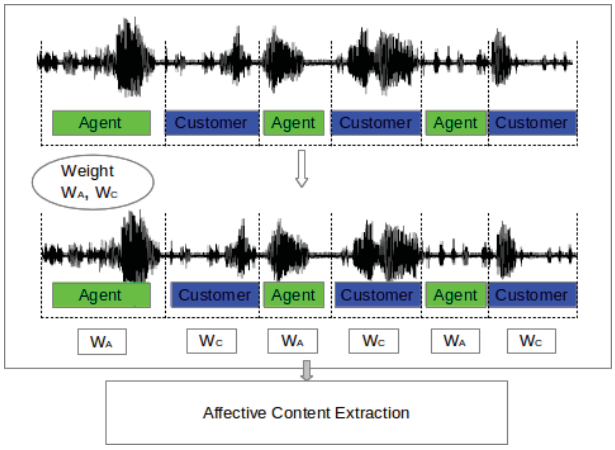

Figure 4: Audio segmentation based knowledge

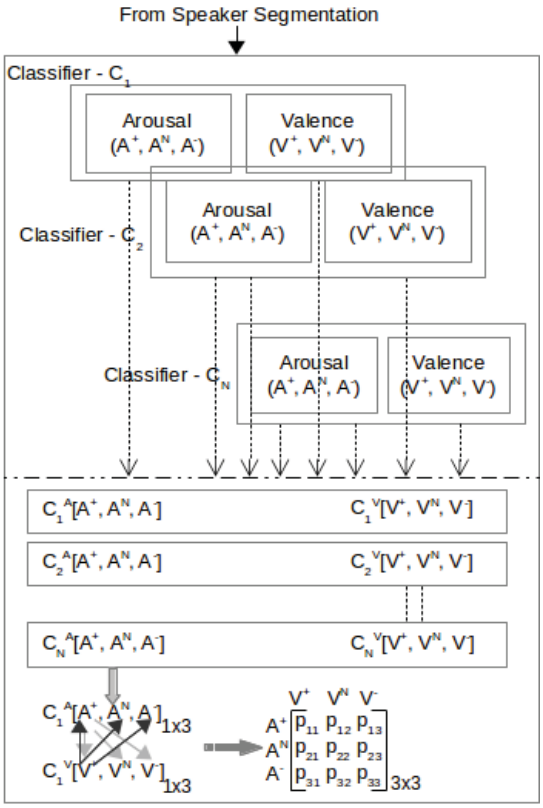

Figure 5: Affective content extraction

larities are measured by calculating the distances between the modified matrix and the reference matrices. In this way, similar segments are bucketed and it becomes easier for the supervisors to analyze the problematic calls depending on the labeled affective contents.

3 Mining similar audio segments

The proposed framework for mining similar audio segments consists of several blocks as shown in Figure 3. The blocks are (i) affective content extractor that gives scores for different affective states in the two dimensional (*arousal* (A) and *valence* (V)) space, (ii) three knowledge-based systems (audio segmentation, time lapse of the segment in the call, and voice to text analytics) (iii) decision block for deciding similar segments. As shown in Figure 3, when an audio call is fed to the framework for the analysis, it is passed through a speaker segmentation system which segments an audio call into different segments of agent's and user's voice. Call center calls consist of voice of the two speakers (agent and customer). Knowledge regarding the speaker (whether an agent or a customer) becomes useful while analyzing a continuous audio call (see Figure 4). As an example, super-agent might be interested to analyze of the customer's voice, not the agent's speech, then there is no need to process the agent's speech. On the other way, if super-agent has to check the performance of agents in terms of their expressed emotions (normal or elicited) in handling calls, the agent's voice are required only for analysis, not the customer's voice. As depicted in the Figure 4, depending on the requirement, the framework is able to pick up all the segments either from the agents or from the customers by selecting appropriate value to (w_A,w_c) pairs. The system choose (1,0) for agent's audio and (0,1) for customer's audio.

3.1 Affective content extraction

Let us assume that we have annotated data from a large corpus in two dimensions, namely A and V. Further, let there be three classes in each affective dimension, namely $E_A = \{A^+, A^N, A^-\}$ and $E_V = \{V^+, V^N, V^-\}$. Let us further assume that we have statistical models for A^+,A^N,A^-, V^+,V^N, and V^- such that we can compute for an audio segment S the following, $P(E_A|S)$ (namely, $P(A^+|S)$, $P(A^N|S)$, and $P(A^-|S)$) and $P(E_V|S)$ (namely, $P(V^+|S)$, $P(V^N|S)$, and $P(V^-|S)$). Therefore, we represent an audio segment S in the A-V space by a 3x3 matrix at the output of each classifier ($1 \leqslant k \leqslant C$),

$$\begin{Bmatrix} P(A^+|S)P(V^+|S) & P(A^+|S)P(V^N|S) & P(A^+|S)P(V^-|S) \\ P(A^N|S)P(V^+|S) & P(A^N|S)P(V^N|S) & P(A^N|S)P(V^-|S) \\ P(A^-|S)P(V^+|S) & P(A^-|S)P(V^N|S) & P(A^-|S)P(V^-|S) \end{Bmatrix}$$

Affective content of a segment S is defined by

$$\epsilon^k_{A,V} = P(E|S) = P(E_A, E_V|S) \qquad (1)$$

where $\epsilon^k_{A,V} = P(E_A, E_V|S)$ is the posterior score associated with S being labeled as emotion E_A and E_V, using a trained recognition system. The posterior can be represented as,

$$P(E_A, E_V|S) = \frac{P(S|E_A, E_V)P(E_A)P(E_V)}{P(S)}$$

where $P(S|E_A, E_V)$ is the likelihood, $P(E_A)$ and $P(E_V)$ are the priors. Assuming that affective contents at the *arousal* and *valence* dimensions are independent, we can write,

$$P(S|E_A, E_V) \approx P(S|E_A)P(S|E_V)$$

Note that S is defined as $\chi(x(\tau - \Delta\tau), x(\tau))$, where χ is the operator that extracts high level features from the audio signal between the time interval $(\tau - \Delta\tau)$ and τ. High level features are the statistical functionals and are constructed from the low level descriptors, which operates in the interval of $x(\tau - \Delta\tau)$ and $x(\tau)$ of the signal. From the output of each classifier, we construct a 3x3 matrix, whose elements are posterior probabilities ($\epsilon^k_{A,V}$) (as shown in Figure 5). According to the Equation 1, $\epsilon^k_{A,V}$ can have two set of elements, ϵ^k_A and ϵ^k_V respectively. As an example, lets say for the first set of classifiers, we have the scores $\epsilon^1_{A^+}$, $\epsilon^1_{A^N}$, and $\epsilon^1_{A^-}$ for the emotions in *arousal* scale. Similarly for the same (first) set of classifier, we have the scores $\epsilon^1_{V^+}$, $\epsilon^1_{V^N}$, and $\epsilon^1_{V^-}$ for the emotions in *valence* scale. Then $\epsilon^k_{A,V}$ (3x3 matrix) is obtained considering each pair from ϵ^k_A and ϵ^k_V. Then the recognition system outputs a posterior probability matrix for the utterance $x(\Delta\tau)$ by combining the scores from all the classifiers, which is given by

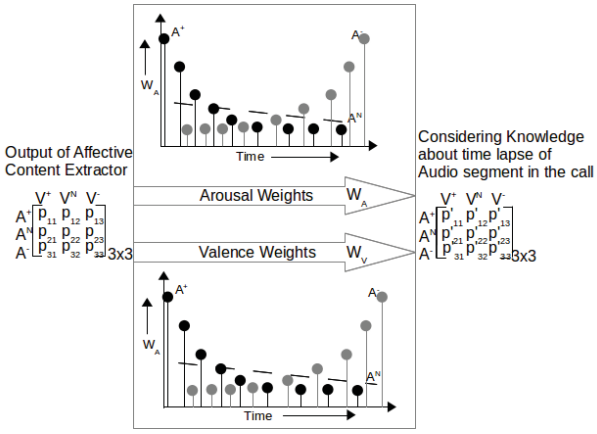

Figure 6: Knowledge regarding the time lapse of the segment in a call

$$\epsilon^{ER}(x(\Delta\tau)) = \sum_{k=1}^{C} \epsilon^k_{A,V} \qquad (2)$$

where $\epsilon^k \in \mathcal{E}_{A,V}$ is the estimated joint probability scores of the utterance $(x(\tau - \Delta\tau)$,$x(\tau))$. This works well with acted speech where one has the luxury of annotated training data (namely, $(x(\Delta\tau), E_A E_V)$ pairs) to build classifiers. However, with the spontaneous speech like call-center calls, probability estimation may be erroneous. However, the estimations can be improved by using knowledge.

3.2 Knowledge about the time lapse of the audio segment in the call

The output probability matrices that we get from affective content extractor is passed through a knowledge-based system, which modifies the probability scores depending upon the time lapse of the segment in the audio call (as shown in Figure 6). We observed through analysis that the duration of the audio calls plays an important role in the induction (or change) in the user's affective state. In such

scenario, we hypothesize that intensity of some of the affective states (namely, A^+, V^-) increases and some (namely, A^-, V^+) decreases. This hypothesis is valid only if no other events occur and change the affective state suddenly. The output can be represented as,

$$\epsilon_i^{lapse} = w_i \epsilon^{ER} \quad (3)$$

where ϵ^{lapse} is the output matrix we get after multiplying weight matrix w_i at the time index i. As shown in the Figure 6 , it is expected that weight values for (A^+,V^-) affective content close to the end of the call will be more in comparison to the beginning of the call. We hypothesize these weight components are expected to increase exponentially as time index i increases for A^+ and V^-, and decrease exponentially as time index i increases for A^- and V^+ (see Figure 6).

3.3 Speech to text analytics

The modified matrices are then passed through the last knowledge-based system (as shown in Figure 7), which converts the spoken utterances into the text format (by using an ASR), followed by the text analytics to generate a weight matrix w_t that consists of the probabilities of affective state given a spoken word or phrase. To analyze the affective state of customer's voice at any instant of time, just previous spoken words of the agent is considered. The same process is followed for analyzing the agent's voice. The hypothesis is that the spoken words from one speaker at any instant of time induce some specific affective state in the other user during a call conversation. This knowledge-based system consists of two sub-system, i) an ASR engine, and ii) a voice analytics. The ASR engine converts the spoken words in the textual format. And the voice analytics system learns and spots emotionally prominent words so as to improve the recognition of emotions. We consider affectively prominent word in audio segment with respect to an affective state is one which appears more often in that category than in other categories of affective states. Like in (Lee and Narayanan, 2005), we also used the prominence measure to find and associate words that are related to affective states in the speech data. With the affective prominence, we create the weight matrix w_t, where each element represent the affective prominence corresponds to each affective state. For each utterance, we have the measure of the affective prominence for each affective states, which forms the weight matrix weight matrix w_t to modify the output of second knowledge-based system (as shown in Figure 7).

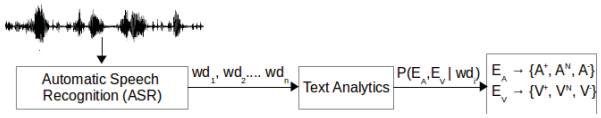

Figure 7: Speech to text analytics

< 0 - No Agreement 0.0 – 0.19 - Poor agreement 0.20 - 0.39 - Fair agreement 0.40 – 0.59 - Moderate agreement 0.60 – 0.79 - Substantial agreement 0.80 – 1.00 - Almost perfect agreement		Kappa Score – Without any knowledge						Kappa Score – Using Knowledge (time Lapse of utterance + linguistic Content)					
		A^+	A^N	A^-	*Raters*	P_i	P_e	A^+	A^N	A^-	*Raters*	P_i	P_e
No of samples	U1	7	0	0	7	1.00	0.47	7	0	0	7	1.00	0.86
	U2	1	1	5	7	0.48		0	1	6	7	0.71	
	U3	3	0	4	7	0.43		7	0	0	7	1.00	
	U4	2	0	5	7	0.52		2	0	5	7	0.52	
	U5	2	0	5	7	0.52		0	0	7	7	1.00	
	U6	5	0	2	7	0.52		6	0	1	7	0.71	
	U7	4	0	3	7	0.43		0	0	7	7	1.00	
	U8	2	2	3	7	0.24		0	7	0	7	1.00	
	U9	2	1	4	7	0.33		7	0	0	7	1.00	
	U10	2	3	2	7	0.24		2	0	5	7	0.52	
	U11	5	0	2	7	0.52		1	0	6	7	0.71	
	U12	0	5	2	7	0.52		7	0	0	7	1.00	
	U13	1	2	4	7	0.33		0	0	7	7	1.00	
	Total	36	14	41	91	**6.10**		39	8	44	91	11.19	
	P_j	0.40	0.15	0.45		$K=(P_i-P_e)/(1-P_e)$		0.43	0.09	0.48		$K=(P_i-P_e)/(1-P_e)$	
	P_e	0.38				$K=$	***0.14***	0.43				$K=$	***0.76***

Figure 8: Kappa statistics in *arousal* dimension with and without knowledge

$$\epsilon^{vta} = w_t \epsilon^{lapse} \quad (4)$$

3.4 Decision based on similarity measures

We get the output matrix ϵ^{vta} for each audio segment and then find the distances from the reference matrix (i.e. the template for each affective state) to calculate the similarity measures. The distance d between ϵ^{vta} and $\epsilon_{A,V}$ are calculated as,

$$d = \| \epsilon^{vta} - \epsilon_{A,V} \| \quad (5)$$

Then the affective state is hypothesized as the output corresponds to the reference matrix that has the minimum distance from the output matrix. It is also possible to arrange the segments in the descending order, i.e. from the highest to the lowest distance. Therefore at the output, audio segment is labeled with the affective class, and its distances from all affective classes.

4 Experiments

4.1 Database

To validate our proposal, we considered call center conversations between the agents and the customers.

Table 1: Affective content detection accuracies for call center calls (%)

Description	Classifiers							
	SVM		ANN		*k*-NN		SVM + ANN + *k*-NN	
	Full Utterance	*400ms Split*	*Full Utterance*	*400ms Split*	*Full Utterance*	*400ms Split*	*Full Utterance*	*400ms Split*
Affective Content Extractor	32.3	62.7	36.1	63.8	31.3	63.2	39.8	65.8
+Time Lapse	49.3	78.3	53.9	78.9	44.2	72.5	59.7	81.9
+ASR and text analytics	56.8	80.9	56.2	82.1	45.1	76.1	61.2	85.2
+Time Lapse + ASR and text analytics	65.3	87.6	61.2	88.9	47	85.6	72.1	89.6

However the agents are trained to behave (and talk) normally in any given situation, how much adverse that might be, mostly suppressing their emotions while talking with the customers. On the other side, customers generally express their emotions while talking to the agents, which means that the customer speech are non-acted (natural). We have considered total 107 call center calls from three different sectors (37 calls from finance, 34 calls from telecommunication sector, and 36 calls from insurance sector). There are total 354 randomly selected audio utterances of the customer which are considered for testing our framework. Notice that each of the spoken utterance has a reference in the form of when in the call flow it was spoken, plus also the manual transcription of speech (temporal sequence of words and phrases transcribed along the duration of the calls). We asked 7 human evaluators to annotate the emotion expressed in each of the 354 utterances by assigning it an emotion label from the set of *arousal* (positive, neutral, and negative) and *valence* (positive, neutral, and negative). In the first set of experiments, we randomly sequenced the utterances (with some utterances repeated) so that the evaluators had no knowledge of the events preceding the audio and we then asked the evaluators to label the utterances; while in the second set of experiments, we provided the utterances in the order in which they were spoken along with the spoken words transcription. The motivation is to include knowledge related to the conversation because of our observation that there is significant disagreement amongst human annotators when they annotate call center speech; the disagreement largely reduces when they are provided with additional knowledge related to the conversation. We computed the Kappa score (Viera and Garrett, 2005) on the annotations in *arousal* dimension for each of the two settings. In the first set of experiments (refer Figure 8), we obtained a score of 0.14 (i.e. without knowledge), suggesting a very poor agreement between the evaluators. While in the second set of experiments we obtained a Kappa score of 0.76 (refer Figure 8) suggesting that there was fair degree of agreement between the evaluators (i.e. with the knowledge). This clearly demonstrates that there was a better consistency in the evaluator's annotation when they were equipped with prior information (knowledge) associated with the utterance. This observation form the basis for the proposed framework for reliable recognition of affective states in call center speech.

4.2 Experimental Results

We considered audio calls which are manually speaker segmented so that the segmentation errors do not propagate to the affective content extractor. Similarly for the speech to text conversion, instead of taking the ASR output, we considered the manual transcription of the audio calls. Affective content extraction system is trained with the acted speech utterances from EmoDB (Emo-DB, 2010). Since the EmoDB dataset has annotations in categorical space, we converted the labels into the dimensional space. All the audio samples in our experimentations are sampled at 8 kHz, 16 bit, and monaural. A low level descriptors (intensity, loudness, 12 MFCC, pitch, voicing probability, F0 envelope, 8 Line Spectral Frequencies, zero-crossing) followed by statistical functionals (maximum, minimum values, range, arithmetic mean, 2 linear regression coefficients, linear and quadratic error, standard deviation, skewness, kurtosis, quartile 1-3, and 3 inter-quartile ranges) are extracted as the meaningful and informative feature sets from each of the segment using the OpenSmile feature extraction toolkit (openSMILE, 2014).

Different classifiers SVM, artificial neural network (ANN), and k-NN have been used in the experiments. LibSVM toolkit is used for implementing SVM classifier (LibSVM, 2015). For ANN, we have

used feed-forward multilayer perceptron (WEKA-Toolkit, 2015), and the network is trained with back-propagation algorithm. All the results are presented as an average detection accuracies over all classes. Two different approaches were adopted for extracting the features from the speech utterances, (1) considering the full utterances (2) splitting the audio in 400 ms like in (Pandharipande and Kopparapu, 2015). In the second case, classifier scores are combined to get the scores for the full utterance.

Table 1 represents the affective content detection accuracies for the segments using different classifiers (and their combination), and using different knowledge-based system. It is observed that combining classifier scores using add rule improves the recognition accuracies (Kuncheva, 2004). The recognition accuracies are improved by using only the knowledge of the segment's lapse in the audio call. Similar trend is observed when only the voice to text analytics knowledge is used, and the accuracies were better compared to the system when only time lapse based knowledge is used. Moreover, the better accuracies are obtained when all the knowledge are incorporated, and the best accuracies are obtained with the framework that segments the full utterance into 400 ms smaller segments compared to the system which uses full utterance for processing. A significant absolute improvement in accuracy of 23.8% is achieved when all the knowledge are used for the combined classifier, and full utterances were segmented into 400 ms smaller segments. We found an average SNR of 8.15 dB for call center calls.

Table 2 presents the affective content detection accuracies for the acted speech samples (from EmoDB dataset), which are contaminated by different levels (SNR level of -5dB to 20dB) of 4 different types of noise (babble, F-16, machine-gun, and volvo) from Noisex-92 dataset. The noise were added using FaNT Toolkit (Filtering and Tool, 2015). Noise contaminated acted speech samples (i.e utterances) are segmented in 400 ms smaller segments like we did in (Pandharipande and Kopparapu, 2015). As expected, for lower SNRs the accuracies are quite on the lower side, and improved with higher SNRs. Combining classifier scores improves the accuracy. Performance of the system degrades significantly when the signal is affected by babble noise, and comparatively lower degradation is observed with the machine-gun noise. However comparing results in Table 1 and 2, we can say that the knowledge-based information significantly helps in improving the performance of the system, even for spontaneous call center calls in real-life noisy environment.

Table 2: Detection accuracies for acted speech (EmoDB) contaminated by noise (Noisex-92)

		Classifiers			
Noise type	**SNR (dB)**	**SVM**	**ANN**	***k*-NN**	**SVM+ANN+*k*-NN**
Babble	-5	21.05	22.3	32.6	33.2
	0	22.1	22.8	32.9	33.9
	5	24.8	25.6	34.6	35.5
	10	28.9	30.1	35.2	37.2
	20	42.3	45.6	47.3	53.3
F-16	-5	20.8	20.8	28.3	30.6
	0	21.6	22.3	29.1	32.4
	5	22.7	23.8	30.5	39.6
	10	28.3	30.1	34.6	43.7
	20	30.7	33.4	38.6	45.2
Machine Gun	-5	22.8	34.6	41.1	47.1
	0	45.6	61.4	61.4	73.3
	5	70.8	71.2	63.2	75.4
	10	71.9	73.2	64.9	77.2
	20	72.3	76	68	80.2
Volvo	-5	20.3	22.8	32.9	37.3
	0	40.3	42.7	40.3	52.9
	5	49.1	49.8	42.1	57.6
	10	54.3	54.8	57.3	71.3
	20	66.7	66.9	72.7	77.3

5 Conclusions

In this paper, we propose a framework that provides an automatic way to extract the affective contents in audio segments of large call center audio calls in *arousal* and *valence* dimension. The system not only relies on the classifier trained with the available acted emotional speech samples, but also incorporates available knowledge related the speech utterances for reliable detection of the affective content. Thus the system provides the call center supervisors an easier way to identify and subsequently address the abnormal calls. Experimental validation suggests that the incorporation of the associated knowledge in terms of speaker information, time lapse of the segment in the call, and linguistic content has improved the performance of the system to reliably identify the affective states and to tag the similar emotional segments. This provides an efficient and useful way for identifying the problematic calls from a large set of recorded call center audio.

References

Emo-DB. 2010. http://www.emodb.bilderbar.info/.

Filtering and Noise Adding Tool. 2015. http://dnt.kr.hs-niederrhein.de/index964b.html?option=com_content&view=article&id=22&Itemid=15&lang=de.

Purnima Gupta and Nitendra Rajput. 2007. Two-stream emotion recognition for call center monitoring. In *INTERSPEECH*.

Sunil Kumar Kopparapu. 2015. *Non-Linguistic Analysis of Call Center Conversations*. Springer, India.

Ludmila I. Kuncheva. 2004. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience.

C M Lee and S S Narayanan. 2005. Towards detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13:293–303.

LibSVM. 2015. https://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Bing Liu. 2012. Sentiment analysis and opinion mining.

Taniya Mishra and Dimitrios Dimitriadis. 2013. Incremental emotion recognition. In *INTERSPEECH*.

M. A. Nicolaou, H. Gunes, and M. Pantic. 2011. Output-associative rvm regression for dimensional and continuous emotion prediction. In *FG*, pages 16–23.

openSMILE. 2014. http://www.audeering.com/research/opensmile.

M. A. Pandharipande and S. K. Kopparapu. 2015. Audio segmentation based approach for improved emotion recognition. In *TENCON 2015 - 2015 IEEE Region 10 Conference*, pages 1–4, Nov.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, jan.

D. Pappas, I. Androutsopoulos, and H. Papageorgiou. 2015. Anger detection in call center dialogues. In *CogInfoCom*.

V. Petrushin. 1999. Emotion in speech: Recognition and application to call centers. In *Artificial Neural Networks in Engineering (ANNIE)*, pages 7–10.

Santiago Planet and Ignasi Iriondo Sanz. 2011. Improving spontaneous children's emotion recognition by acoustic feature selection and feature-level fusion of acoustic and linguistic parameters. In *Advances in Nonlinear Speech Processing - 5th International Conference on Nonlinear Speech Processing, NOLISP 2011, Las Palmas de Gran Canaria, Spain, November 7-9, 2011. Proceedings*, pages 88–95.

Laurence Vidrascu and Laurence Devillers. 2007. Five emotion classes detection in real-world call center data the use of various types of paralinguistic features. In *PARALING*, pages 11–16.

A. J. Viera and J. M. Garrett. 2005. Understanding interobserver agreement: the kappa statistic. *Family Medicine*, 37(5):360–363.

WEKA-Toolkit. 2015. http://www.cs.waikato.ac.nz/ml/weka/.