Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 96 (2016) 587 – 596

Procedia Computer Science

20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

Knowledge-based framework for intelligent emotion recognition in spontaneous speech

Rupayan Chakraborty*, Meghna Pandharipande, Sunil Kumar Kopparapu

TCS Innovation Labs - Mumbai, Yantra Park, Thane(West)-400 601, INDIA

Abstract

Automatic speech emotion recognition plays an important role in intelligent human computer interaction. Identifying emotion in natural, day to day, spontaneous conversational speech is difficult because most often the emotion expressed by the speaker are not necessarily as prominent as in acted speech. In this paper, we propose a novel spontaneous speech emotion recognition framework that makes use of the available knowledge. The framework is motivated by the observation that there is significant disagreement amongst human annotators when they annotate spontaneous speech; the disagreement largely reduces when they are provided with additional knowledge related to the conversation. The proposed framework makes use of the contexts (derived from linguistic contents) and the knowledge regarding the time lapse of the spoken utterances in the context of an audio call to reliably recognize the current emotion of the speaker in spontaneous audio conversations. Our experimental results demonstrate that there is a significant improvement in the performance of spontaneous speech emotion recognition using the proposed framework.

Peer-review under responsibility of KES International

Keywords: Knowledge-based framework; emotion recognition; intelligent systems; spontaneous speech; non-acted emotion

* Corresponding author. Tel.: +91-22-6778-8295 ; fax: +91-22-6778-1581.
E-mail address: rupayan.chakraborty@tcs.com, meghna.pandharipande@tcs.com, sunilkumar.kopparapu@tcs.com

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2016.08.239

1. Introduction

Human emotions play an important role in intelligent human computer interactions. Much of the initial emotion recognition research has been successfully validated on acted speech (for example [1,2,3,4]). With the introduction of call centers associated with the growing services industry, the focus has shifted to spontaneous speech [1] [5,6,7,8,9,10,11]. Speech emotion recognition systems that perform with high accuracies on acted speech datasets do not perform as well on realistic natural speech [12]. This can be attributed to the mismatch in train-test datasets, however the fact remains that acted speech is an exaggeration of emotions which is not a characteristic in spontaneous speech. There are two problems associated with spontaneous speech, namely (i) building a spontaneous speech database suitable for emotion recognition and (ii) reliable emotion annotation of spontaneous speech by human annotators. In spite of these problems, emotion recognition of spontaneous natural speech has attracted the attention of researchers (for example [13,8,12,14,15,16,17,18]). In [14], authors proposed a combination of three sources of information (i.e. acoustic, lexical, and discourse) for emotion recognition in spoken dialogue system and found improvements in recognition performance. In [19], authors described an approach to improve emotion recognition in spontaneous children's speech by combining acoustic and linguistic features. More recently, with a view to identify emotions in near real time, an incremental emotion recognition system has been proposed that updates the recognized emotion with each recognized word in the conversation [20]. They make use of three features (i.e. cepstral, intonation and textual), obtained at the word level to estimate the emotion with better accuracies.

In this paper, we propose a framework for emotion recognition that can work for both spontaneous and acted speech. The framework is a combination of several modules, each of which extracts information related to the emotion. Combining the output of these modules produces a better estimate of the emotion. Unlike [20], we do not rely only 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. Our framework is motivated by the hypothesis that the emotion in a spontaneous speech utterance at any instance of time not only depends on instantaneously extracted emotion, but is also dependent on (a) time lapse of the utterance in the audio call, and (b) context-based information (events derived from linguistic contents). We validate our proposed framework in different and diversified scenarios of both acted and spontaneous speech through experiments. Moreover, the usefulness of knowledge incorporation is tested with two spontaneous datasets in two different train-test conditions, where the emotion models are generated from (a) acted speech samples (mismatched scenario) (b) spontaneous speech samples (matched scenario). The framework and its validation in realistic scenarios are the main contributions of this paper.

The rest of the paper is organized as follows. Section 2 presents the challenges in determining emotion in spontaneous speech and motivates the proposed framework. In Section 3, we propose the framework for emotion recognition that incorporates knowledge-based information. Section 4 describes the datasets, experiments and results. We conclude in Section 5.

2. Motivation

Our motivation leading to the framework for emotion recognition of spontaneous speech is based on two observations, namely, (a) *human perception for emotion is better in acted speech compared to spontaneous speech since the former exhibits higher discriminating characteristics than the latter* [12], and (b) *human annotators do better on spontaneous speech when they are given the context associated with the speech that they are annotating* [14,8,21]. According to the first observation, Figure 1 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 exhibits higher degree of intensity, both in *arousal* and *valence* dimensions resulting in a larger radii emotion vector compared to the spontaneous speech. For this reason, it is easy to mis-recognize one emotion for another in spontaneous speech. Subsequently, if the first quadrant (Figure 1) represents emotion E_1 and the fourth quadrant represents emotion E_2, then the mis-recognition error is small (Δr) for spontaneous speech but

[1] We will use the word spontaneous, non-acted and natural speech interchangeably in this paper

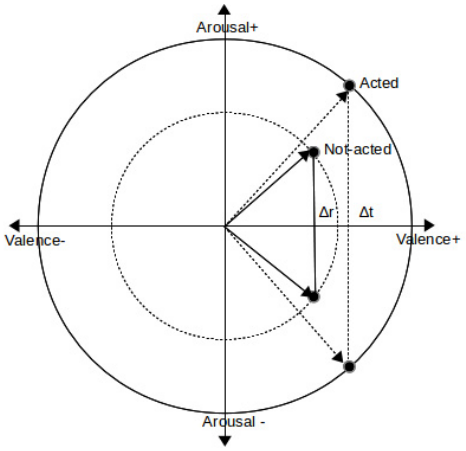

Fig. 1: Error in emotion estimation (acted vs spontaneous (non-acted))

Agreements:- < 0 - No agreement 0.0 - 0.19 - Poor 0.20 - 0.39 - Fair 0.40 - 0.59 - Moderate 0.60 - 0.79 - Substantial 0.80 – 1.00 - Almost perfect		Kappa Score – Without any knowledge							Kappa Score – Using Knowledge (time Lapse of utterance + linguistic Content)						
		A	H	N	S	Raters	p_i	p_e	A	H	N	S	Raters	p_i	p_e
No of samples	U1	7	0	0	0	7	1.00	0.40	7	0	0	0	7	1.00	0.79
	U2	1	1	5	0	7	0.48		0	1	6	0	7	0.71	
	U3	3	0	4	0	7	0.43		0	0	7	0	7	1.00	
	U4	2	0	4	1	7	0.33		2	0	5	0	7	0.52	
	U5	2	0	4	1	7	0.33		0	0	7	0	7	1.00	
	U6	5	0	1	1	7	0.48		6	0	1	0	7	0.71	
	U7	3	0	3	1	7	0.29		0	0	6	1	7	0.71	
	U8	1	2	4	0	7	0.33		0	1	6	0	7	0.71	
	U9	2	1	3	1	7	0.19		0	0	7	0	7	1.00	
	U10	2	1	1	3	7	0.19		2	0	0	5	7	0.52	
	U11	5	0	1	1	7	0.48		1	0	0	6	7	0.71	
	U12	0	0	2	5	7	0.52		0	0	1	6	7	0.71	
	U13	1	2	3	1	7	0.19		0	0	7	0	7	1.00	
	Total	34	7	15	35	91	5.24		18	2	18	53	91	10.33	
	p_j	0.37	0.08	0.16	0.38	$K = (p_o - p_e)/(1-p_e)$			0.198	0.022	0.198	0.582	$K = (p_o - p_e)/(1-p_e)$		
	p_e	0.32				**K =**	**0.12**		0.42				**K=**	**0.65**	

Fig. 2: Kappa score: annotators agreement in an IVR-SERES call

requires higher degree of error in judgment (Δt) to mis-recognize emotion E_1 as emotion E_2 and vice-versa for acted speech.

Regarding the second observation, we experienced that there is a fair amount of disagreement among the annotators when they are not provided any knowledge related to spoken utterances. Let us explain it for a call, which is taken from the spontaneous dataset of Interactive Voice Response (IVR) Speech Enabled Railway Enquiry System (SERES)[22,23] (as shown in Figure 2). The call consists of 13 spoken utterances (U1-U13), and we asked 7 human evaluators to annotate the emotion expressed in each of the utterances by assigning it an emotion label from the set {*anger*, *happy*, *neutral*, *sad*}. In the first set of experiments, we randomly sequenced the utterances so that the evaluators had no knowledge of the events preceding the audio and then the evaluators were asked to label (with scores between 1-5) the utterances. While in the second set of experiments, we provided the utterances in the order in which they were spoken in response to the voice prompts of the IVR system (as shown in Figure 2).

We computed the Fleiss' Kappa score for each of the two settings[24,25]. In the first set of experiments, we obtained a score of 0.12 (see Figure 2), suggesting a very poor agreement between the evaluators. While in the second set of experiments, we obtained a Kappa score of 0.65 suggesting that there was fair degree of agreement between the evaluators. This clearly demonstrates that there was a better consistency in the evaluator's annotation when they were

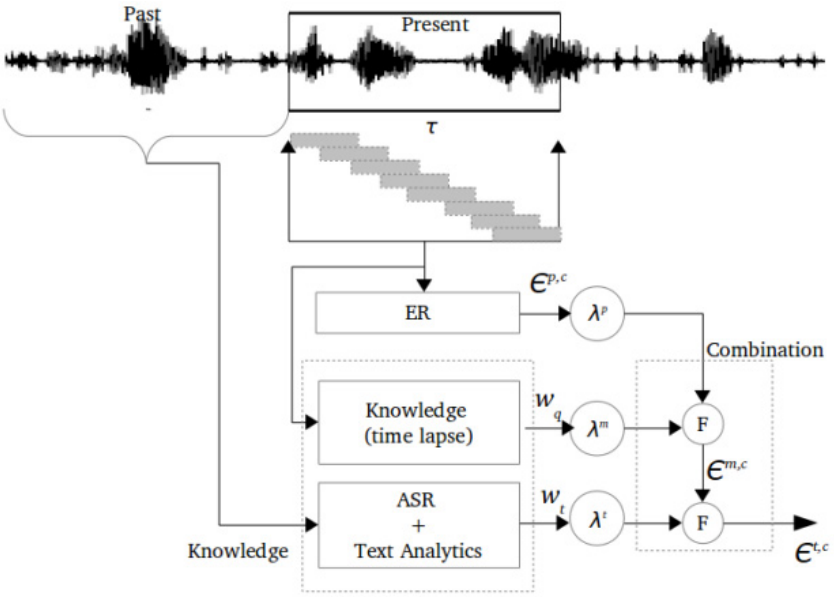

Fig. 3: Knowledge-based framework for emotion recognition

equipped with prior information (knowledge) associated with the spoken utterance. These two observations form the basis for the proposed framework for recognizing emotions in spontaneous speech.

3. Framework for spontaneous speech emotion recognition

Human evaluators are more consistent annotating spontaneous speech when they have prior knowledge about the spoken utterance. We believe this is primarily because the intensity of emotion displayed in spontaneous speech is small compared to the intensity displayed in acted speech. The proposed framework tries to address these aspects (as depicted in Figure 3). Let $x(\tau)$ be the speech utterances whose emotion is to be determined, and let there be n possible emotions, namely $\mathcal{E} = (E_1 = anger, E_2 = happy, \cdots, E_n)$ associated with any speech utterance. Emotion at any point of time τ for a classifier ($1 \leq c \leq C$) is defined by

$$\epsilon_k^p = P(E_k|x(\tau)) = \frac{P(x(\tau)|E_k)P(E_k)}{P(x(\tau))} \tag{1}$$

where $\epsilon_k^p = P(E_k|x(\tau))$ is the posterior score associated with $x(\tau)$ being labeled as emotion E_k using some trained emotion recognition system. Note that $x(\tau)$ is represented as $\chi(x(\tau))$, where χ is the operator that extracts the relevant features from the whole utterance of time interval τ.

Conventionally, the emotion of the utterance $x(\tau)$ is given by

$$E_{k*}(x(\tau)) = \arg\max_{1\leq k\leq n}\{\epsilon_k^p\} \tag{2}$$

where $E_{k*} \in \mathcal{E}$ is the estimated emotion of the utterance.

3.1. Knowledge about the time lapse of speech utterance in the call

As depicted in Figure 3 and 4, the output scores from emotion recognizer are given as input to a knowledge-based system, that modifies the scores depending upon the time lapse of the utterances in the audio signal. We observed that the duration of the audio calls plays an important role in the induction (or change) in the user's emotion. Therefore, we hypothesize that the intensity of some emotions (namely, *anger* and *sad*) increases and some (namely, *happy*)

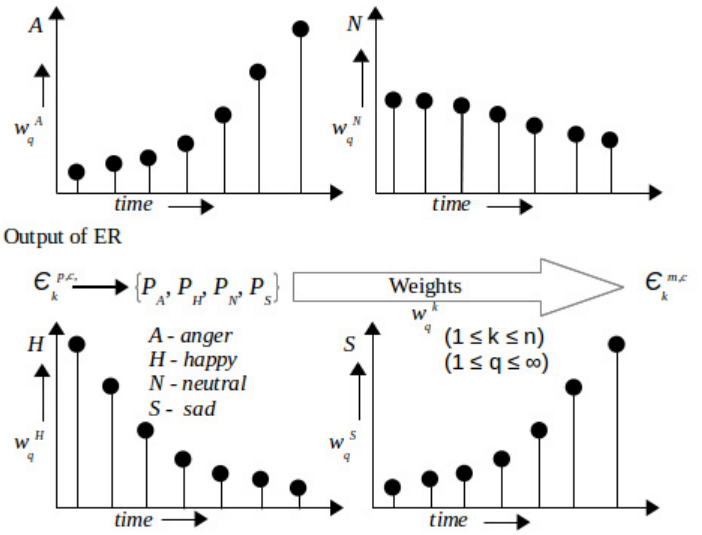

Fig. 4: Knowledge regarding the time lapse of the utterances in the call

decreases. This hypothesis is valid only if no other events occur and change the emotion suddenly. The effect of this can be represented as,

$$\epsilon_k^m = w_q \epsilon_{k,q}^p \tag{3}$$

where $\epsilon_{k,q}^p$ and w_q are the probability score and weight vector at time instant q (see Figure 4). It is expected that the influence of previous events close to the present will be more compared to those which are further away from the present. We hypothesize that w_q is expected to increase or decay exponentially as q increases, depending upon the type of the emotion and its change in a specific application case. As an example and as shown in the Figure 4, it is expected that weight values for *anger* and *sad* 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 increases for *anger* and *sad*, and decrease exponentially as time index increases for *happy* (see Figure 4).

3.2. Linguistic contents and event-based knowledge

The scores ϵ_k^m are then fed to the knowledge-based system that converts the spoken utterances into the text format (by using an ASR), followed by the text analytics to generate a weight vector w_t, which consists of the probabilities of emotion given the spoken words or phrase (refer Figure 5). To analyze the emotion at any instant of time, previous spoken words are considered. It takes into account the prior events (or contexts) that are related to the linguistic contents. The hypothesis is that the spoken words from one speaker at any instant of time induce some specific emotion in the other user during the call conversation. Here, two tasks are performed, i) speech to text conversion by ASR and ii) voice analysis (learns and spots emotionally prominent words so as to improve the recognition of emotions). Emotionally prominent words (or phrases) in audio utterances with respect to an emotion is one which appears more often in that category than in other categories of emotion. We used the prominence measure to find and associate words that are related to emotions in the speech data. With the emotional prominence, we create the weight matrix w_t, where each element represent the emotional prominence corresponds to each emotion. In calculating emotional prominence, we denote the words (or phrases) in the utterances by W=(wd_1,wd_2,...wd_j) and the set of emotions of the utterances by E_k, then the self mutual information is given by,

$$i(wd_j, E_k) = \log \frac{P(E_k|wd_j)}{P(E_k)} \tag{4}$$

where $P(E_k|wd_j)$ is the posterior probability that an utterance containing words wd_j implies emotion class E_k. It is understandable that if the word wd_j in an utterance highly correlates to an emotion, then $P(E_k|wd_j) > P(E_k)$, and

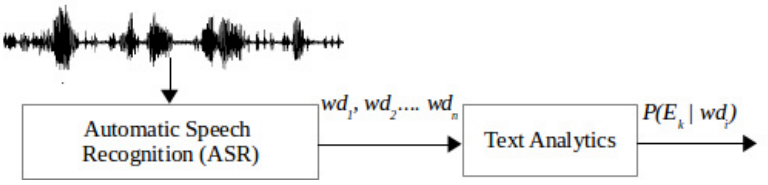

Fig. 5: Linguistic content-based knowledge

$i(wd_j, E_k)$ is positive. And if, the word wd_j in an utterance is not correlated to an emotion, then $P(E_k|wd_j) < P(E_k)$, and $i(wd_j, E_k)$ is negative. If there is no effect by the word, then $i(wd_j, E_k) = 0$. The emotional prominence $prom(wd_j)$ of a word for an emotion is defined as mutual information between a specific word and emotion, and is given by,

$$prom(wd_j) = \sum_{j=1}^{M} P(E_k|wd_j)i(wd_j, E_k) \tag{5}$$

Therefore, emotional prominence is a measure of the amount of information that a specific word contains about a given emotion category. For each spoken utterance, the weight vector w_t (contains both linguistic and context-based information) modifies the previous system's output ϵ_k^m. It can be expressed as

$$\epsilon_k^t = w_t \epsilon_k^m \tag{6}$$

As will be shown in our experimental results, we can combine ϵ_k^p, ϵ_k^m, and ϵ_k^t to *better* estimate the emotion of the spontaneous utterance $x(\tau)$. We can generalize this as

$$e_k^c = F\left\{\lambda_p \epsilon_k^p, \lambda_m \epsilon_k^m, \lambda_t \epsilon_k^t\right\} \tag{7}$$

where F is a combining function (in all our experiments, we assumed F to be an addition operator), and λs are the binary weights.

Emotion of the spontaneous speech utterance $x(\tau)$ with the incorporation of knowledge (ϵ_k^m and ϵ_k^t) is represented as

$$E_{k*} = \arg\max_{1\le k\le n}\left\{\sum_{c=1}^{C} e_k^c\right\} \tag{8}$$

where $\sum_{c=1}^{C} e_k^c$ represents the combination of scores from multiple classifiers [26].

Knowledge regarding the time lapse of the utterance in an audio call and the linguistic contents, especially in conversational system, provides useful information to recognize the emotion of the speaker. Therefore, incorporation of these knowledge could be useful in extracting the exact emotion of an user.

Note that when $\lambda_m = \lambda_t = 0$ in (7), the framework boils down to the conventional method used to compute emotion in literature. We conjecture that, while computing emotion with $\lambda_p = 1, \lambda_m = \lambda_t = 0$ might be useful for acted speech, it is far from sufficient if one has to work with spontaneous speech.

4. Experiments

Experiments were conducted on both spontaneous speech datasets [23,27] and acted dataset (EmoDB [28]) to check the validity of the proposed framework. Detailed description of the spontaneous dataset is provided in Table 1. Using Fleiss' Kappa statistics (which is used for multiple labelers) [24,25]; we obtain kappa score "k" in the range of 0.68-0.70, which corresponds to substantial inter-labeler agreement when the annotators were provided with the additional information about the calls. Utterances having kappa score $k \geq 0.6$ suggests substantial agreement are used here in experimentation. The number of utterances are 1264 and 2117 respectively for two datasets (namely, IVR-SERES and Call center). Training-testing sample distribution is always kept (80-20)% ratio for all experiments reported here.

Table 1: Spontaneous speech dataset details

	IVR-SERES	CALL-CENTER
Domain	Telephonic	Telephonic
Total no. of calls	117	210
Total Duration (min)	308	850
Avg no. of user utterances(turn) per call	18	27
No. of annotators	7	7
Kappa score without knowledge	0.14	0.17
Using substantial agreement by Kappa statistics		
Kappa score with knowledge	0.68	0.70
Total no. of utterances in experimentation	1264	2117
No. of utterances in training set	1012	1694
No. of utterances in testing set	253	424
Avg utterances length (sec)	3.1	11

	Probability Score			
Phrases/Words	**A**	**H**	**N**	**S**
I apologize	0.3	0	0.1	0.6
Thank you so much	0	0.8	0.2	0
outstanding pending	0.4	0	0.3	0.3
advised replacement	0.5	0	0.2	0.3
I transfer your call	0.4	0	0.2	0.3
oh my its about	0.5	0	0.5	0

Fig. 6: Emotional prominence for spoken utterances in a call

We performed data balancing so that the classifiers are trained uniformly for all classes. All the audio samples in our experimentations are sampled at 8 kHz, 16 bit, and monoaural.

OpenSMILE paralinguistic 384 dimension audio features (earlier used for Interspeech 2009 Emotion Challenge[29]) are extracted and reduced to a lower dimension after feature selection using WEKA Toolkit[30]. Different classifiers SVM, artificial neural network (ANN), and k-NN have been used in the experiments. LibSVM toolkit is used for implementing SVM classifier[31]. For SVM, we used polynomial kernel and pairwise multi-class discrimination based on Sequential Minimal Optimization. ANN was trained using Levenberg-Marquardt backpropagation algorithm[32], and having three layers: input, output, and hidden. Number of neurons used in the hidden layer was selected through extensive experimentations. To be fair in our comparison, we reported results in all our experimentations for 4 emotion classes: *anger*, *happy*, *sad*, *neutral*(most frequently observed emotions in spontaneous call conversations[5,9]).

Table 2 shows the performance of the emotion recognition system on spontaneous (IVR-SERES and Call center) and acted (EmoDB) speech datasets using 3 different classifiers, and their combination. While testing the system on spontaneous datasets, we conducted two sets of experiments. In the first, emotion samples from the acted EmoDB dataset were used to train the classifier, thus resulting a train-test mis-matched condition (represented as "U" in Table 2). We checked how the associated knowledge improve the spontaneous speech emotion recognition performance even if the classifiers are trained with the samples from a different dataset (i.e. acted in this case). This situation may arise if someone does not have the luxury of using annotated data because of the fact that annotation requires substantial amount of human intervention and cost; and may not be available for call center calls because of the infrastructure issues. Most often call center calls contain confidential information regarding customers, which may not be disclosed due to legal issues. Conversely, in the second set of experiment, emotion samples from the same datasets were used to train the classifier ("M" in Table 2 to represent "*matched*" condition). As can be seen, SVM classifier performs consistently best among all the classifiers. The classifier combination further improves the recognition accuracy.

To generate the result by relying only on the output from emotion recognizer and not incorporating any knowledge, λ_p is set to 1, while λ_m and λ_t are kept 0 (presented in first rows of Table 2 for three datasets). Note that, w_q and w_t vectors are learnt separately from the available meta-data associated to the calls (transcriptions and associated time stamps of the spoken utterances), which are used for training. The vector w_q is learnt from the training dataset

by averaging the scores given by the human annotators for each emotion class. For longer test calls, vector size is expanded through padding by the value of the last element of the vector that was obtained during training. Vector "w_t" related to emotional prominence is learnt from the available manual text transcription of the calls (used in training). We observed that people tend to use specific word for expressing their emotions. In fact, while listening to the data that was used to tag the emotion classes, the annotators reported that they felt some specific emotions if they heard certain words in the utterances. Figure 6 shows the values of emotional prominence for some utterances (i.e. words or phrase) from the training data. It is natural that people tend to use certain words more frequently in expressing their emotions because they learned the correlation between certain words and the related emotions [33].

Table 2: Emotion recognition accuracies (%) on spontaneous and acted datasets

IVR - SERES (Spontaneous speech)											
				SVM		**ANN**		**KNN**		**SVM+ANN+KNN**	
Description	λ_p	λ_m	λ_t	*U*	*M*	*U*	*M*	*U*	*M*	*U*	*M*
conventional(baseline), ϵ^p	1	0	0	35.3	68.1	37.8	68.9	36.8	59.8	40.2	71.3
+time lapse, ϵ^m	1	1	0	44.3	70.2	48	73.1	37	62.9	52.1	72.9
+ linguistic contents, ϵ^t	1	0	1	51.7	73.6	57.8	75.6	38.1	64	61.4	79.1
+time lapse + linguistic content	1	1	1	56.1	77.1	61.9	76.3	49.6	67.3	67.1	82.1
CALL CENTER (Spontaneous speech)											
				SVM		**ANN**		**KNN**		**SVM+ANN+KNN**	
Description	λ_p	λ_m	λ_t	*U*	*M*	*U*	*M*	*U*	*M*	*U*	*M*
conventional(baseline), ϵ^p	1	0	0	35.2	66.7	36.1	65.2	31.3	60.3	39.8	69.8
+time lapse, ϵ^m	1	1	0	49.1	72.3	53.9	73	39.2	66.1	56.7	74.2
+ linguistic contents, ϵ^t	1	0	1	49.8	72.7	49.2	72.8	37.1	64.8	53.2	76.9
+time lapse + linguistic content	1	1	1	55.3	76	51.2	73.6	45	68.9	65.1	78.1
EMODB (Acted speech)											
				SVM		**ANN**		**KNN**		**SVM+ANN+KNN**	
Description	λ_p	λ_m	λ_t	*M*		*M*		*M*		*M*	
conventional(baseline), ϵ^p	1	0	0	83.4		80.8		78.3		87.3	

As can be seen in Table 2, the system that relies only on ϵ^p (i.e. when λ_p=1 and λ_m=λ_t=0) gives a low performance accuracy for spontaneous speech dataset (best accuracy of 40.2% obtained by combining classifier for mismatched condition). This is to be expected, since the recognizer is trained with the acted speech samples, but tested with the spontaneous speech (a mismatch in training-testing). The accuracy improves by an absolute value of 31.1% in matched scenario. However, the results improve when the knowledge related to either time lapse of the utterance in the call ($\lambda_p, \lambda_m \neq 0$) or the linguistic content ($\lambda_p, \lambda_t \neq 0$) is used. Finally, when both of these knowledge are used (namely $\lambda_p, \lambda_m, \lambda_t \neq 0$), system produces the best recognition accuracies for mismatched condition; an absolute improvements of 26.9% and 25.3% respectively are achieved for two spontaneous datasets (IVR-SERES and call center). The best accuracies of 82.1% (for IVR-SERES) and 78.2% (for Call Center) are achieved for matched scenario. Interestingly, for IVR-SERES dataset, more improvement was found when linguistic information are used in comparison to the time lapse related information. However, the improvement is small for the call center calls while time lapse related information is used. We hypothesize that this could be because of the nature of the datasets; the utterances in IVR setup are more *isolated* compared to the utterances in a call center setup which leads to time lapse based knowledge less contributing to the improvement in emotion recognition accuracies. We also observed that the performance accuracy on acted dataset are better compared to the spontaneous dataset. The classifiers were better learnt with acted speech, even with the small number of samples per class in comparison to the spontaneous speech. This might be due to the acted samples were recorded in the controlled environment with minimum background noise, whereas the spontaneous speech utterances (especially from users) are from different backgrounds of having variable types and levels of noise. For speech to text conversion, instead of using the ASR output we considered the manual transcription for the spoken utterances to ensure that ASR errors do not propagate through the system.

Figure 7 shows a test IVR-SERES call and the recognized emotions using the knowledge-based framework. It validates the usefulness of the inclusion of linguistic context-based knowledge (namely, ϵ_k^t). First column represents the manual transcription of the audio call conversation between an user and IVR computer (represented as "system" in Figure 7). Second column represents the scores (for 4 classes) at the output of the emotion recognition system, and the

	Scores using ϵ^p					Scores using ϵ^l					Combined $\epsilon^p + \epsilon^l$
IVR Call	k=1 (A)	k=2 (H)	k=3 (N)	k=4 (S)	Emotion	k=1 (A)	k=2 (H)	k=3 (N)	k=4 (S)		Emotion (scores)
System: welcome to the system											
System: please select a language											
User: English	0.1	0.1	0.8	0.0	N	-	-	-	-	-	N (0.8)
System: which functionality would you select?											
User: seat availability	0.1	0.0	0.7	0.1	N	-	-	-	-	-	N(0.7)
System: unable to understand	Event Occurred										
User: seat availability	**0.2**	**0.5**	**0.2**	**0.1**	**H**	**1.0**	**0.0**	**0.0**	**0.0**	**A**	**A(1.2)**
System: which train?											
User: train (x)	0.6	0.1	0.1	0.2	A	-	-	-	-	-	A(0.6)
System: which station (from)?											
User: station (A)	0.4	0.0	0.6	0.0	N	-	-	-	-	-	N(0.6)
System: which station (to)?											
User: station (B)	0.2	0.1	0.7	0.0	N	-	-	-	-	-	N(0.7)

Fig. 7: Linguistic context improves emotion recognition in a IVR-SERES call

fourth column represents the vectors at the output of linguistic knowledge-based system. The last column represents the maximum score obtained (i.e. obtained by combining the scores presented in second and fourth column) and its corresponding class label. It is very hard to expect the emotion of an user to be *happy* when (s)he has to repeat the same query (*seat availability*) for the second time due to the fact that system is not able to recognize what is spoken by the user (here it is an event or context). Our framework allows for correction the emotion by using the event (or context) based knowledge ($\lambda_p, \lambda_t \neq 0$). The conventional system (namely, ϵ^p) outputs *happy*, whereas after combining the event-based knowledge (namely, ϵ^l), the output becomes *anger*.

5. Conclusions

Recognizing emotion in spontaneous speech is difficult because it does not carry sufficient intensity to distinguish one emotion from the other. In this paper, we hypothesize that the lack of discriminating properties in the audio can be handled by making use of prior knowledge in the form of linguistic contents and time lapse of the utterances in the call. The main contribution of this paper is the development of a framework for spontaneous speech emotion recognition. The framework is generic in the sense that it reduces to the conventional method of emotion recognition when $\lambda_m = \lambda_t = 0$. The experimental results validate the use of different context dependent knowledge in terms of the time lapse of the utterances in the audio call and the linguistic contents. The framework has been evaluated on three different databases in both matched and unmatched train-test conditions. Incorporation of prior knowledge and classifier fusion to better the emotion recognition accuracy for spontaneous speech, even in an unmatched train-test scenario, clearly establishes the usefulness of the proposed framework.

References

1. Schuller, B.W., Batliner, A., Steidl, S., Seppi, D.. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication* 2011;**53**:1062–1087.
2. Ayadi, M.E., Kamel, M.S., Karray, F.. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition* 2011;**44**:572–587.
3. Mower, E., Mataric, M., Narayanan, S.S.. A framework for automatic human emotion classification using emotion profiles. *IEEE Transactions on Audio, Speech and Language Processing* 2011;**19**(5):1057–1070.
4. Wu, S., Falk, T.H., Chan, W.Y.. Automatic speech emotion recognition using modulation spectral features. *Speech Communication* 2010; **53**(5):768–785.
5. Petrushin, V.. Emotion in speech: Recognition and application to call centers. In: *Artificial Neural Networks in Engineering (ANNIE)*. 1999, p. 7–10.
6. Burkhardt, F., van Ballegooy, M., Englert, R., Huber, R.. An emotion-aware voice portal. *Proc Electronic Speech Signal Processing ESSP* 2005;:123–131.
7. Vidrascu, L., Devillers, L.. Five emotion classes detection in real-world call center data the use of various types of paralinguistic features. In: *PARALING*. 2007, p. 11–16.
8. Burkhardt, F., Polzehl, T., Stegmann, J., Metze, F., Huber, R.. Detecting real life anger. In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2009, p. 4761–4764. doi:10.1109/ICASSP.2009.4960695.
9. Gupta, P., Rajput, N.. Two-stream emotion recognition for call center monitoring. In: *INTERSPEECH*. 2007, .
10. Kopparapu, S.K.. *Non-Linguistic Analysis of Call Center Conversations*. SpringerBriefs in Electrical and Computer Engineering. Springer International Publishing; 2014. ISBN 9783319008974. URL: http://books.google.co.in/books?id=Uew1BAAAQBAJ.
11. Pappas, D., Androutsopoulos, I., Papageorgiou, H.. Anger detection in call center dialogues. In: *Cognitive Infocommunications (CogInfoCom), 2015 6th IEEE International Conference on*. 2015, p. 139–144. doi:10.1109/CogInfoCom.2015.7390579.
12. Schuller, B.W., Seppi, D., Batliner, A., Maier, A.K., Steidl, S.. Towards more reality in the recognition of emotional speech. In: *ICASSP*. 2007, p. 941–944.
13. Yacoub, S., Simske, S., Lin, X., Burns, J.. Recognition of emotions in interactive voice response systems. In: *In: Proc. of Eurospeech*. 2003, p. 729–732.
14. Lee, C.M., Narayanan, S.S.. Towards detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing* 2005; **13**:293–303.
15. Schuller, B., Müeller, R., Höernler, B., Höethker, A., Konosu, H., Rigoll, G.. Audiovisual recognition of spontaneous interest within conversations. In: *ICMI*. New York, NY, USA: ACM. ISBN 978-1-59593-817-6; 2007, p. 30–37. URL: http://doi.acm.org/10.1145/1322192.1322201. doi:10.1145/1322192.1322201.
16. Tarasov, A., Delany, S.J.. Benchmarking classification models for emotion recognition in natural speech: A multi-corporal study. In: *FG*. 2011, p. 474–477.
17. Polzehl, T., Schitt, A., Metze, F.. *Spoken Dialogue Systems Technology and Design*; chap. Salient Features for Anger Recognition in German and English IVR Portals. New York, NY: Springer New York. ISBN 978-1-4419-7934-6; 2011, p. 83–105. URL: http://dx.doi.org/10.1007/978-1-4419-7934-6_4. doi:10.1007/978-1-4419-7934-6_4.
18. Neiberg, D., Elenius, K., Laskowski, K.. Emotion recognition in spontaneous speech using gmms. In: *INTERSPEECH*. 2007, p. 809–812.
19. Planet, S., Iriondo, I.. *Improving Spontaneous Children's Emotion Recognition by Acoustic Feature Selection and Feature-Level Fusion of Acoustic and Linguistic Parameters*. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-25020-0; 2011, p. 88–95. URL: http://dx.doi.org/10.1007/978-3-642-25020-0_12". doi:10.1007/978-3-642-25020-0_12.
20. Mishra, T., Dimitriadis, D.. Incremental emotion recognition. In: *INTERSPEECH 2013*. 2013, p. 2876–2880.
21. Steidl, S., Levit, M., Batliner, A., Noth, E., Niemann, H.. Of all things the measure is man : Automatic classification of emotions and inter-labeler consistency. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*; vol. 1. 2005, p. 317–320. doi:10.1109/ICASSP.2005.1415114.
22. Bhat, C., Mithun, B.S., Saxena, V., Kulkarni, V., Kopparapu, S.. Deploying usable speech enabled ivr systems for mass use. In: *Human Computer Interactions (ICHCI), 2013 International Conference on*. 2013, p. 1–5.
23. AWAZYP-SERES, . Speech Enabled Railway Enquiry System. https://www.youtube.com/watch?v=fSaNk8CZtGY&feature=youtu.be; 2014.
24. Viera, A.J., Garrett, J.M.. Understanding interobserver agreement: the kappa statistic. *Family Medicine* 2005;**37**(5):360–363.
25. McHugh, M.L.. Interrater reliability: the kappa statistic. *Biochemia Medica* 2012;**3**:276–282.
26. Kuncheva, L.I.. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience; 2004. ISBN 0471210781.
27. AWAZYP-CallCenter, . Non-linguistic Analysis of Call Center Call Conversations. https://sites.google.com/site/sunilkopparapu/Home/nonlinguistic-analysis-of-call-center-conversation; 2014.
28. Emo-DB, . Berlin Database of Emotional Speech. http://www.emodb.bilderbar.info/; 2013.
29. Schuller, B.W., Steidl, S., Batliner, A.. The INTERSPEECH 2009 emotion challenge. In: *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*. 2009, p. 312–315.
30. WEKA-Toolkit, . http://www.cs.waikato.ac.nz/ml/weka/; 2015.
31. LibSVM, . A Library for Support Vector Machines. https://www.csie.ntu.edu.tw/~cjlin/libsvm/; 2015. [Online; accessed December-2015].
32. Haykin, S.. *Neural Networks: A Comprehensive Foundation*. Upper Saddle River, NJ, USA: Prentice Hall PTR; 2nd ed.; 1998. ISBN 0132733501.
33. Plutchik, R.. *The Psychology and Biology of Emotion*. Comparative Government S. HarperCollinsCollegePublishers; 1994. ISBN 9780060452360. URL: https://books.google.co.in/books?id=9O5-AAAAMAAJ.