Chapter 3
Do You Mean What You Say? Recognizing Emotions in Spontaneous Speech

Rupayan Chakraborty, Meghna Pandharipande and Sunil Kumar Kopparapu

Abstract Emotions when explicitly demonstrated by an actor are easy for a machine to recognize by analyzing their speech. However in case of day to day, naturally spoken spontaneous speech it is not easy for machines to identify the expressed emotion even though emotion of the speaker are embedded in their speech. One of the main reasons for this is that people, especially non-actors, do not explicitly demonstrate their emotion when they speak, thus making it difficult to recognize the emotion embedded in their spoken speech. In this paper, based on some of our previous published work (example, Chakraborty et al. in Proceedings of the 20th International Conference KES-2016 96:587–596, 2016 [1], Chakraborty et al. in TENCON 2015—2015 IEEE Region 10 Conference 1–5, 2015 [2], Chakraborty et al. in PACLIC, 2016 [3], Pandharipande and Kopparapu in TENCON 2015—2015 IEEE Region 10 Conference 1–4, 2015 [4], Kopparapu in Non-Linguistic Analysis of Call Center Conversations, 2014 [5], Pandharipande and Kopparapu in ECTI Trans Comput Inf Technol 7(2):146–155, 2013 [6], Chakraborty and Kopparapu in 2016 IEEE International Conference on Multimedia and Expo Workshops, 1–6, 2016 [7]) we identify the challenges in recognizing emotions in spontaneous speech and suggest a framework that can assist in determining the emotions expressed in spontaneous speech.

Keywords Spontaneous speech · Emotion · Recognition · Classification · Challenges

R. Chakraborty · M. Pandharipande · S.K. Kopparapu (✉)
TCS Innovation Labs, Mumbai, India
e-mail: SunilKumar.Kopparapu@TCS.Com

R. Chakraborty
e-mail: Rupayan.Chakraborty@TCS.Com

M. Pandharipande
e-mail: Meghna.Pandharipande@TCS.Com

S.R.S. Prabaharan et al. (eds.), *Frontiers in Electronic Technologies*,
Lecture Notes in Electrical Engineering 433, DOI 10.1007/978-981-10-4235-5_3

1 Introduction

Several nuances are embedded in human speech. A spoken utterance can be analyzed for *what* was spoken (speech recognition), *how* was it spoken (emotion recognition) and *who* spoke (speech biometric) it. Most often these three aspects form the basis of most of the work being carried out actively by speech researchers. In this paper, we concentrate on the *how* aspect of spoken utterance, namely emotion recognition.

Perceiving emotions from different real-life signals is a natural and an inherent characteristic of a human being. For this reason emotion plays a very important role in intelligent human–computer interactions. Machine perception of human emotion not only helps machine to communicate more humanely, but it also helps in improving the performance of other associated technologies like Automatic Speech Recognition (ASR) and Speaker Identification (SI).

With the mushrooming of services industry there has been a significant growth in the voice-based call centers (VbCC) where identifying emotion in spoken speech has gained importance. The primary goal of a VbCC is to maintain a high level of customer satisfaction which means understanding the customer just in time (in real time and automatically) and making a decision on how to communicate (what to say, how to say) with the customer. While several things related to the customer are a priori available, thanks to the advances in data mining, the one thing that is crucial is the emotion of the customer at that point of time, so that the agent can plan what and how to converse to keep the customer happy and also allow him to know when to make a pitch to up-sell.

Much of the initial emotion recognition research has been successfully validated on acted speech (for example [8–11]). Emotions expressed by trained actors are easy to recognize, primarily because they are explicitly and dramatically expressed by them with significant intensity. This, on purpose magnified, emotions can be easily distinguished from one another. However, when the expression of the emotion is not explicit or loud, it is very difficult to distinguish one emotion of the speaker from another. This mild and not explicitly demonstrated emotion is most likely to occur in spontaneous natural day to day conversational speech. The rest of the paper is organized as follows. In Sect. 2 we dwell on the different challenges facing emotion recognition in spontaneous speech. We propose a framework in Sect. 3 that has provision to use prior knowledge to address emotion recognition in spontaneous speech. And we conclude in Sect. 4.

2 Challenges

Historically emotion has been represented using two affective dimensions, namely, *arousal* (also referred to as activation) and *valence*. Note that any point in this 2D space (Fig. 1) can be looked upon as a vector and represents an emotion. Table 1 gives the mapping of a few emotions in terms of the affective states. For

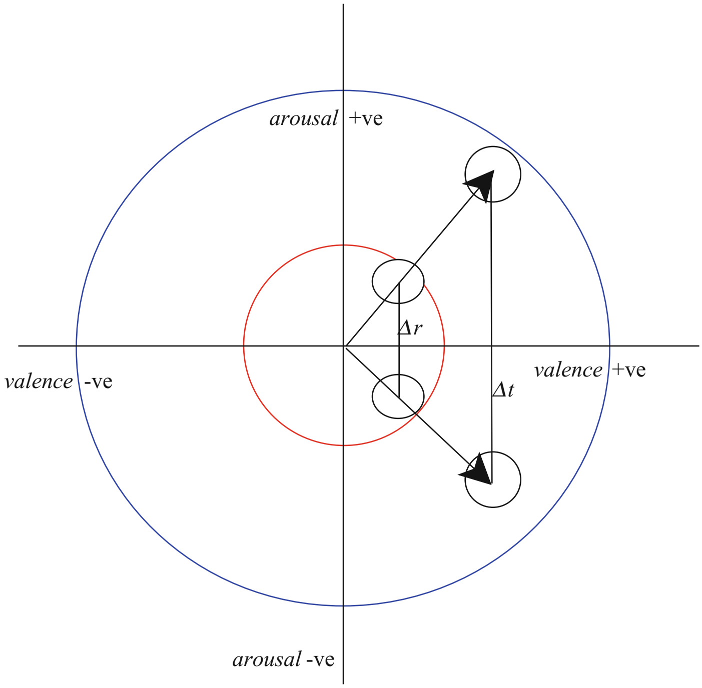


Fig. 1 Emotions expressed in the (*arousal*, *valence*) space. Emotion in spontaneous speech is subtle compared to acted speech

Table 1 Emotions expressed in the (*arousal*, *valence*) space. Map of the affective states to known human emotion

arousal	*valence*	Emotion
+	+	*happy*
+	−	*anger*
0	0	*neutral*
−	−	*sad*

example +ve *valence* and +ve *arousal* (first quadrant) would represent *happy* while −ve *valence* and +ve *arousal* could represent *anger*. Now we enumerate the challenges in machine recognizing emotion in spontaneous speech.

Intensity of Emotion in Spontaneous Speech Usually acted speech exhibits higher degree of intensity, both in *arousal* and *valence* dimensions resulting in a larger radii emotion vector compared to the spontaneous (non-acted) speech. For this reason, it is easy to mis-recognize one emotion for another in spontaneous speech. Subsequently, if the first quadrant (Fig. 1) represents emotion E_1 and the fourth quadrant represents emotion E_2, then the misrecognition error is small (Δr) for spontaneous 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. For this reason, recognizing emotion in spontaneous speech becomes challenging and is more prone to misrecognition.

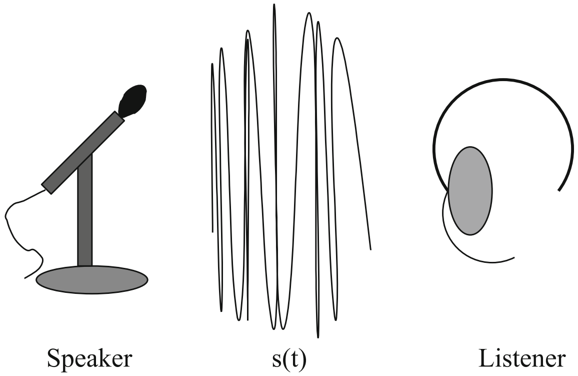
What works for acted speech does not work for Spontaneous Speech Though recognizing emotion in speech has a rich literature, however, most of the work has been done on acted speech are typically machine learning-based systems. Namely, one trains a system (example, Support Vector Machine, Artificial Neural Networks, Deep Neural Networks) with a set of annotated speech data and then classifies the test dataset using the trained system. Speech emotion recognition systems that perform with high accuracies on acted speech datasets do not perform as well on realistic natural speech [12] because of the mismatch between the train (acted) and test (spontaneous) datasets. This is another challenge in addressing spontaneous speech emotion recognition.

Clearly this challenge can be addressed if there exists an emotion annotated spontaneous speech dataset which can be used for training a spontaneous speech emotion recognition system.

Spontaneous Speech Corpus For any given spoken utterance, there are two point of views in terms of associating emotion to the utterance namely, (a) encoded emotion and (b) decoded emotion. The emotional state of the speaker who uttered the audio is called the encoded emotion while the interpreted emotion of the same audio by a listener, who is different from the speaker, is called decoded emotion. For example, the audio $s(t)$ in Fig. 2 can have two emotion labels associated with it. When the speaker annotates and assigns a emotion label it is called the encoded emotion of $s(t)$ and when a listener (different from the speaker) listens to $s(t)$ and assigns an emotion label it is called the decoded emotion.

For acted speech both decoded and encoded emotion are more likely to be the same, however, for spontaneous speech there is bound to be a wide gap between the encoded and decoded emotion. Building a realistic spontaneous speech corpus would need a person to speak in a certain emotional state and/or annotate what

Fig. 2 Decoded versus encoded emotions



he or she spoke; generating such realistic spontaneous data corpus is extremely difficult and is a huge challenge.

Annotating Spontaneous Speech Corpus The next best thing is to have a decoded spontaneous speech corpus. However, one of the problems associated with emotion recognition of spontaneous speech, is the availability of a reliably emotion annotated spontaneous speech database suitable for emotion recognition. The inability to annotate spontaneous speech corpus is basically because of the lower degree of emotion expression (as seen in Fig. 1).

In [1] we showed that there is a fair amount of disagreement among the evaluators when they are asked to annotate spontaneous spoken utterances. The disagreement, however, decreases when the evaluators are provided with the context knowledge associated with the utterance. Fleiss' Kappa score [13, 14] was used to determine the agreement between evaluators. When the evaluators were asked to annotate (decoded emotion) spontaneous speech the agreement was 0.12 while the same set of evaluators when provided with the context associated with the spontaneous speech, the agreement between the evaluators increased to 0.65. This suggests that there is a higher degree of agreement between the evaluators when they are provided associated contextual knowledge while annotating spontaneous speech.

As illustrated above, there are several known challenges that exist in spontaneous speech emotion recognition. Clearly the literature that deals with emotion recognition of acted speech does not help in spontaneous speech emotion recognition, however, as observed, the use of prior knowledge can help address recognizing emotions in spontaneous speech. In the next section we propose a framework for recognizing emotions in spontaneous speech based on this observation.

3 A Framework for Spontaneous Speech Emotion Recognition

Let $s(t)$ be a speech signal, say of duration T seconds and let

$$\mathscr{E} = (E_1 = anger, E_2 = happy, \dots, E_n)$$

be the set of n emotion labels. In literature the emotion of the speech signal $s(t)$ is computed as

$$\mu^p_{k,s(t)} = P(E_k|s(t)) = \frac{P(s(t)|E_k)P(E_k)}{P(s(t))} \quad (1)$$

where $\mu^p_{k,s(t)} = P(E_k|s(t))$ is the posterior probability or score associated with $s(t)$ being labeled as emotion $E_k \in \mathscr{E}$. Generally, these posteriors are calculated by learning the likelihood probablities from a reliable training dataset using some machine learning algorithm. Note that in practice the features extracted from the speech signal

$\mathscr{F}(s(t))$ are used instead of the actual raw speech signal $s(t)$ in (1). Conventionally, the emotion of the speech signal $s(t)$ is given by

$$E_{k*} = \arg \max_{1 \leq k \leq n} \{\mu^{p}_{k,s(t)}\}. \tag{2}$$

Note that $E_{k*} \in \mathscr{E}$ is the estimated emotion of the signal $s(t)$.

While this process of emotion extraction works well for acted speech, because the entire speech utterance carries one emotion. However this, namely the complete speech signal carrying a single emotion is seldom true for spontaneous conversational speech (for example, a call center audio recording between the agent and a customer). As mentioned in an earlier section, additional challenges exists in terms of the fact that emotions in spontaneous speech are not explicitly demonstrated and hence can not be robustly identified even by human annotators in the absence of sufficient context surrounding the spoken utterance.

These observations lead us to look for a novel framework for recognizing emotions in spontaneous speech [1]. The framework tries to take care of the fact that (a) the emotion within the same speech signal is not the same and (b) human annotators are better able to recognize emotions when they are provided with a context associated with the speech signal.

The essential idea of the framework is to compute emotion for smaller duration ($2\Delta\tau$) segments of the speech signal ($s_{\tau}(t)$), namely,

$$s_{\tau}(t) = s(t) \times \{U(\tau - \Delta\tau) - U(\tau + \Delta\tau)\}$$

where $U(t)$ is a unit step function defined as

$$\begin{aligned} U(t) &= 1 \quad \text{for} \quad t \geq 0 \\ &= 0 \quad \text{for} \quad t < 0, \end{aligned}$$

instead of computing the emotion for the complete signal ($s(t)$). Note that (a) $s_{\tau}(t) \subset s(t)$ and is of length $2\Delta\tau$ and (b) $\tau \in [0, T]$. As done conventionally, the emotion of $s_{\tau}(t)$ is computed as earlier, namely,

$$\mu^{p}_{k,s_{\tau}(t)} = P(E_k | s_{\tau}(t)) \tag{3}$$

for $k = 1, 2, \ldots n$. However, in addition we also make use of the emotions computed from the previous η speech segments, namely $\mu_{k,s_{\tau-v}(t)}$ for $v = 1, 2, \ldots \eta$. So we have, the posterior score associated with the speech utterance $s_{\tau}(t)$ being labeled E_k as

$${}'\mu^{p}_{k,s_{\tau}(t)} = \mu^{p}_{k,s_{\tau}(t)} + \sum_{v=1}^{\eta} \omega_v \mu^{p}_{k,s_{\tau-v}(t)} \tag{4}$$

where $\omega_1, \omega_2 \ldots, \omega_{\eta}$ are monotonically decreasing weights, which are all less than 1. Equation (4) makes sure that the posterior score of the speech segment $s_{\tau}(t)$ is influ-

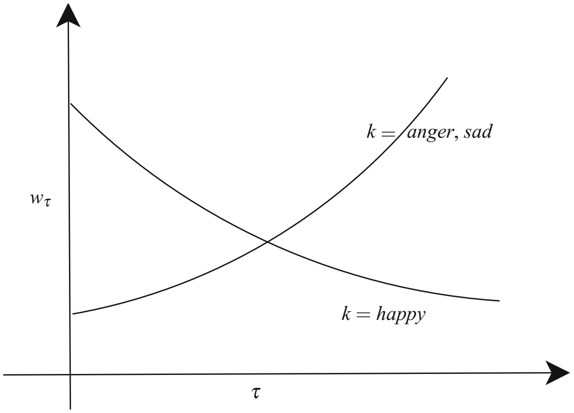


Fig. 3 Knowledge regarding the time lapse of the utterances in the call. The weights w_τ of emotions like *happy* decreases with τ while the weights increase for emotions like *anger*, *sad*

enced by the weighted sum of the posterior score of the previous speech segments. This is generally true of spontaneous conversational speech where the emotion of the speaker is based on the past emotion experienced during the conversation.

Further the output posterior scores from emotion recognizer, namely, $\mu^p_{k,s_\tau(t)}$ (3) is given as input to a knowledge-based system, that modifies the scores depending upon the time lapse (how far is τ from the beginning of the spoken conversation) of the speech segment (utterance) in the audio signal. This can be represented as,

$$\mu^\kappa_{k,s_\tau(t)} = w_\tau \mu^p_{k,s_\tau(t)} \tag{5}$$

where $\mu^p_{k,s_\tau(t)}$ (3) and w_τ (see Fig. 3) are the posterior probability score and weight vector at time instant τ respectively. And $\mu^\kappa_{k,s_\tau(t)}$ is the emotion computed based on knowledge.

The motivation for (5) is based on the observation that the duration of the audio calls plays an important role in the induction (or change) in the user's emotion. As mentioned in [1] weight w_τ is expected to increase or decay exponentially as τ increases, depending upon the type of the emotion. As an example (see Fig. 3) it is expected that w_τ for *anger* and *sad* close to the end of the conversation is likely to be more compared to the same emotion of the customer at the beginning of the call. As seen in Fig. 3 the weight components are expected to increase exponentially as time index increases for *anger* and *sad* at the same time w_τ is expected to decrease exponentially as time index increases for *happy* emotion.

We can combine μ^p_k and μ^κ_k to *better* estimate the emotion of the spontaneous utterance $s_\tau(\tau)$ as

$$e^k = \lambda_p('\mu^p_k) + \lambda_\kappa(\mu^\kappa_k) \tag{6}$$

where $\lambda_\kappa = 1 - \lambda_p$. The framework makes use of knowledge when $\lambda_\kappa \neq 0$. Emotion of the spontaneous speech utterance $s(\tau)$ with the incorporation of knowledge (μ^κ_k) is represented as

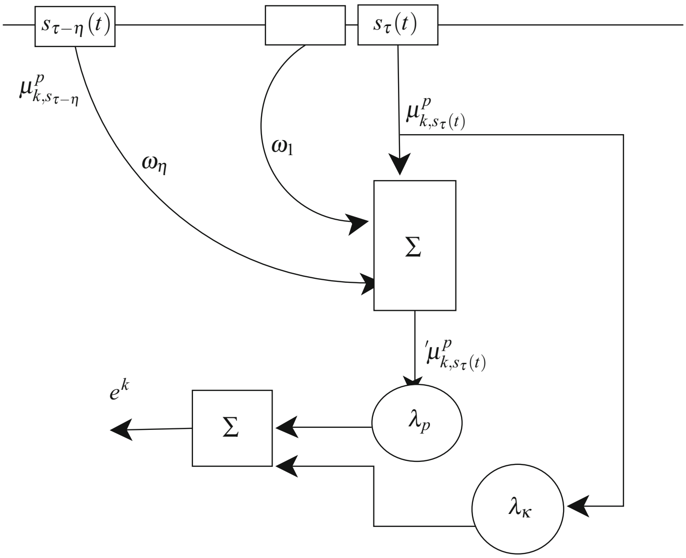


Fig. 4 Proposed framework for spontaneous speech emotion recognition

$$E_{k*} = \arg \max_{1\leq k\leq n} \left\{e^k\right\}. \tag{7}$$

Knowledge regarding the time lapse of the utterance in an audio call, especially in conversational system, provides useful information to recognize the emotion of the speaker. Therefore, incorporation of this knowledge is useful in extracting the actual emotion of an user. The proposed framework for spontaneous speech recognition is shown in Fig. 4.

As shown in [1] there is performance improvement in recognition of emotion of spontaneous speech when this framework is used. They show for different classifiers that there is almost 11% absolute improvement in emotion recognition for interactive voice response type of call and the performance further improved to 14% absolute for real call center conversation.

4 Conclusion

Emotion recognition has rich literature for acted speech and this leads to the belief that the techniques that work well for acted speech can be directly used for spontaneous speech. However, there are several dissimilarities between acted and spontaneous speech which does not allow one to use techniques and algorithms that work well for acted speech to recognize emotion in spontaneous speech. Emotion recognition techniques are generally machine learning based algorithms which (a) requires sufficient amount of training data and (b) requires the test and the train data to match. The main challenge in using trained models that work for acted speech on spontaneous speech is the mismatched condition. Additionally, in case of spontaneous speech it is very challenging to (a) generate spontaneous speech data and (b) to obtain robust annotation of the speech data. For this reason techniques and algorithms that work best for spontaneous speech cannot be built afresh. In this paper, we first established the importance of spontaneous speech emotion recognition and then enumerated several challenge and hurdles faced during emotion recognition in spontaneous speech. Based on our previous work, we proposed a framework that exploits apriori knowledge to enable reliable spontaneous speech emotion recognition. The main idea behind the proposed framework is to assist the machine learning algorithm with prior knowledge associated with the spontaneous speech. It has been shown [1] that this framework can actually improve the emotion recognition accuracies of spontaneous speech by as much as 11–14% in absolute terms.

References

1. R. Chakraborty, M. Pandharipande, S.K. Kopparapu, Knowledge-based framework for intelligent emotion recognition in spontaneous speech, in *Procedia Computer Science, 2016, knowledge-Based and Intelligent Information; Engineering Systems: Proceedings of the 20th International Conference KES-2016*, vol. 96, pp. 587–596. http://www.sciencedirect.com/science/article/pii/S187705091632049X
2. R. Chakraborty, M. Pandharipande, S. Kopparapu, Event based emotion recognition for realistic non-acted speech, in *TENCON 2015—2015 IEEE Region 10 Conference* (2015), pp. 1–5
3. R. Chakraborty, M. Pandharipande, S.K. Kopparapu, Mining call center conversations exhibiting similar affective states, in *PACLIC 2016* (2016)
4. M.A. Pandharipande, S.K. Kopparapu, Audio segmentation based approach for improved emotion recognition, in *TENCON 2015—2015 IEEE Region 10 Conference* (2015), pp. 1–4
5. S.K. Kopparapu, *Non-Linguistic Analysis of Call Center Conversations*, Springer Briefs in Electrical and Computer Engineering (Springer, 2014)
6. M.A. Pandharipande, S.K. Kopparapu, A language independent approach to identify problematic conversations in call centers. ECTI Trans. Comput. Inf. Technol. **7**(2), 146–155 (2013)
7. R. Chakraborty, S.K. Kopparapu, Improved speech emotion recognition using error correcting codes, in *2016 IEEE International Conference on Multimedia and Expo Workshops, ICME Workshops 2016, Seattle, WA, USA, July 11–15, 2016*. IEEE Computer Society (2016), pp. 1–6. doi:10.1109/ICMEW.2016.7574707
8. B.W. Schuller, A. Batliner, S. Steidl, D. Seppi, Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Commun. **53**, 1062–1087 (2011)
9. M.E. Ayadi, M.S. Kamel, F. Karray, Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn. **44**, 572–587 (2011)
10. E. Mower, M. Mataric, S.S. Narayanan, A framework for automatic human emotion classification using emotion profiles. IEEE TASLP **19**(5), 1057–1070 (2011)
11. S. Wu, T.H. Falk, W.Y. Chan, Automatic speech emotion recognition using modulation spectral features. Speech Commun. **53**(5), 768–785 (2010)
12. B.W. Schuller, D. Seppi, A. Batliner, A.K. Maier, S. Steidl, Towards more reality in the recognition of emotional speech, in *ICASSP* (2007), pp. 941–944
13. A.J. Viera, J.M. Garrett, Understanding interobserver agreement: the kappa statistic. Family Med. **37**(5), 360–363 (2005)
14. M.L. McHugh, Interrater reliability: the kappa statistic. Biochemia Med. **3**, 276–282 (2012)