Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 83-89 recibido 04-04-2018 revisado 29-04-2018 aceptado 04-05-2018

Bi-modal annoyance level detection from speech and text

Detección del nivel de enfado mediante un sistema bi-modal basado en habla y texto

Raquel Justo, Jon Irastorza, Saioa Pérez, M. Inés Torres
Universidad del País Vasco UPV/EHU. Sarriena s/n. 48940 Leioa. Spain
{raquel.justo,manes.torres}@ehu.eus

Abstract: The main goal of this work is the identification of emotional hints from speech. Machine learning researchers have analysed sets of acoustic parameters as potential cues for the identification of discrete emotional categories or, alternatively, of the dimensions of emotions. However, the semantic information gathered in the text message associated to its utterance can also provide valuable information that can be helpful for emotion detection. In this work this information is included within the acoustic information leading to a better system performance. Moreover, it is noticeable the use of a corpus that include spontaneous emotions gathered in a realistic environment. It is well known that emotion expression depends not only on cultural factors but also on the individual and on the specific situation. Thus, the conclusions extracted from the present work can be more easily extrapolated to a real system than those obtained from a classical corpus with simulated emotions.

Keywords: speech processing, semantic information, emotion detection on speech, annoyance tracking, machine learning

Resumen: El principal objetivo de este trabajo es la detección de cambios emocionales a partir del habla. Diferentes trabajos basados en aprendizaje automático han analizado conjuntos de parámetros acústicos como potenciales indicadores en la identificación de categorías emocionales discretas o en la identificación dimensional de las emociones. Sin embargo, la información semántica recogida en el mensaje textual asociado a cada intervención, puede proporcionar información valiosa para la detección de emociones. En este trabajo se combina la información textual y acústica dando lugar a mejoras en el rendimiento del sistema. Es importante recalcar por otra parte, el uso de un corpus que incluye emociones espontáneas recogidas en un entorno realista. Es bien sabido que la expresión de la emoción depende no solo de factores culturales si no también de factores individuales y de situaciones particulares. Por lo tanto, las conclusiones extraídas en este trabajo se pueden extrapolar más fácilmente a un sistema real que aquellas obtenidas a partir de un corpus clásico en el que se simula el estado emocional.

Palabras clave: procesamiento del habla, información semántica, reconocimiento emocional en el habla, rastreo del enfado, aprendizaje automático

1 Introduction

The detection of emotional status has been widely studied in the last decade within the machine learning framework. The goal of researchers is to be able to recognise emotional information from the analysis on voice, language, face, gestures or ECG (Devillers, Vidrascu, y Lamel, 2005). One of the main important challenges that need to be faced in this area is the need of supervised data, i.e. corpora including human data annotated with emotional labels (Devillers, Vidrascu, y Lamel, 2005) (Vidrascu y Devillers, 2005) and this is not a straightforward task due to the subjectivity of emotion perception by humans (Devillers, Vidrascu, y Lamel, 2005) (Eskimez et al., 2016). Many works considered corpora that consist of data from professional actors simulating the emotions to be analyzed. However, it usually leads to poor

ISSN 1135-5948. DOI 10.26342/2018-61-9

results due to many factors like the differences among the real situations the detection system has to deal with and the emotional status picked up in the corpus. Moreover, the selection of valuable data including spontaneous emotions depends on the goals of the involved research and it is difficult to find an appropriate corpus that matches the specific goal of each task.

Focussing on emotion identification from speech and language a wide range of potential applications and research objectives can be found (Valstar et al., 2014) (Wang et al., 2015) (Clavel y Callejas, 2016). Some examples are early detection of Alzheimer's disease (Meilán et al., 2014), the detection of valency onsets in medical emergency calls (Vidrascu y Devillers, 2005) or in Stock Exchange Customer Service Centres (Devillers, Vidrascu, y Lamel, 2005). Emotion recognition from speech signals relies on a number of short-term features such as pitch, additional excitation signals due to the non-linear air flow in the vocal tract, vocal tract features such as formants (Wang et al., 2015) (Ververidis y Kotropoulos, 2006), prosodic features (Ben-David et al., 2016) such as pitch loudness, speaking rate, rhythm, voice quality and articulation (Vidrascu y Devillers, 2005) (Girard y Cohn, 2016), latency to speak, pauses (Justo et al., 2014) (Esposito et al., 2016), features derived from energy (Kim y Clements, 2015) as well as feature combinations, etc. Regarding methodology, statistical analysis of feature distributions has been traditionally carried out. Classical classifiers such as the Bayesian or SVM have been proposed for the identification of emotional characteristics from speech signals. The model of continuous affective dimensions is also an emerging challenge when dealing with continuous rating of emotion labelled during real interaction (Mencattini et al., 2016). In this approach recurrent neural networks have been proposed to integrate contextual information and then predict emotion in continuous time to just deal with arousal and valence (Wollmer et al., 2008) (Ringeval et al., 2015).

When regarding text, there are numerous works dealing with sentiment analysis whose application domains range from business to security considering well-being, politics or software engineering (Cambria, 2016). However, there are few works considering the recognition of specific emotions such as joy, love or anger (Medeiros y van der Wal, 2017; Gilbert y Karahalios, 2010; Marsden y Campbell, 2012). Moreover, it seems reasonable to think that the combination of acoustic and textual information might lead to improve emotion recognition systems performance. However, although there are plenty of research articles on audio-visual emotion recognition, only a few research works have been carried out on multimodal emotion recognition using textual clues with visual and audio modality (Eyben et al., 2010; Poria et al., 2016).

In this work we deal with a problem proposed by a Spanish company providing customer assistant services through the telephone (Justo et al., 2014). They want to automatically detect annoyance rates during customer calls for further analysis, which is a novel and challenging goal. Their motivation is to verify if the policies applied by operators to deal with annoyed and angry customers lead to shifts in customer behavior. Thus an automatic procedure to detect those shifts will allow the company to evaluate their policies through the analysis of the recorded audios. Moreover, they were interested in providing this information to the operators during the conversation. As a consequence this work is aimed at detecting different levels of annoyance during real phone-calls to Spanish complain services. Mainly, we wanted to analyse the effect of including textual information into the annoyance detection system based on acoustic signals.

The paper is organised as follows, Sec. 2 describes the previous work carried out to solve the presented problem with the specific dataset we are dealing with. In Sec. 3 the annotation procedure in terms of speech signal and text is described and Sec. 4 details the experiments carried out and the obtained results. Finally, Sec. 5 summarises the concluding remarks and future work.

2 Dataset and previous work

The Spanish call center company involved in this work offers customer assistance for several phone, tv and internet service providers. The customer complaint services of these companies receive a certain number of phone-calls from angry or annoyed customers. But the way of expressing annoyance is not the same for all the customers. Some of them are furious and shout; others speak quickly with frequent and very short micropauses but do not shout (Justo et al., 2014), others seems to be more fed-up than angry; others feel impotent after a number of service failures and calls to the customer service. The dataset for this study consisted of seven conversations between customers and the call-centre operators that were identified and selected by experienced operators. All the selected customers were very angry with the service provider because of unsolved and repeated service failures that caused serious troubles to them. In a second step each recording was named according to the particular way the customer expresses his annoyance degree. Thus, call-center operators qualified the seven subjects in conversations as follows: *Disappointed, Angry (2 records), Extremely angry, Fed-up, Impotent and annoyed in disagreement.* All these feelings correspond to the different ways the customer in the study expressed their annoyance with the service provided. More specifically they correspond to the way the human operators perceived customer feelings. The duration of the conversations was 42s, 42s, 35s, 16m20s, 1m08s, 1m02s and 1m35s respectively, resulting in a total of 22.1 minutes.

In a previous work (Irastorza y Torres, 2016) the different records were manually annotated. Two members of the research group acted as expert annotators. They first identified customer speech segments, agent speech segments and overlapping segments. Only intelligible customer speech segments were considered for the experiments. In a second step, annotators were asked to identify the changes in the degree of perceived emotion in each recording using zero for neutral or very low, one for medium and two for high degree. They were asked to mark time steps where they perceived a change in the degree of expression and then label each segment with the corresponding perceived level. The annotator agreement was high in the identification of the time steps where they perceived changes in the degree of expression. Then, just one of the two annotations was chosen to fix segments bounds. However, the procedure resulted in a significant level of disagreements when regarding the label given to each step. Thus, most frequent disagreements were considered as new levels in the proposed scale of annoyance expression. The resulting set of categories consists of five degrees defined as follows: *very low* agreed by annotators, *low*, which corresponds to a low-medium disagreement, *medium* agreed by annotators, *high*, which corresponds to a medium-high disagreement and *very high* agreed by annotators. Less frequent disagreements were not considered. The right side of Table 1 (SPEECH-BASED) shows the final number of segments identified for each audio file and annoyance level.

An automatic classification was carried out in (Irastorza y Torres, 2016) using acoustic parameters extracted from the audio files. The acoustic signal was divided into 20 ms overlapping windows (frames) from which a set of features was extracted. The classification procedure was carried out over those frames. A combination of Intensity and intensity-based suprasegmental features along with LPC coefficients achieved the highest frame classification accuracies for all the expressions of annoyance analysed. The obtained results validated the annotation procedure and also showed that shifts in customer annoyance rates could be potentially tracked during phone calls.

3 Text annotation process

The transcription of the utterances provides additional information related to the semantics, language style, among others, that cannot be found in acoustics and might help in the detection of annoyance or other emotional categories. Thus, in this work the text associated with the utterances was considered and annotated to be included in the classification process.

3.1 Transcription

The audio files described in the previous section were transcribed in order to use text as an additional information source. The transcriptions were carried out making use of *Praat* (Boersma y Weenink, 2016) software tool. The segments obtained from the aforementioned annotation, in terms of time steps, were also employed here and only those intelligible customer speech segments were transcribed.

In this case, given that there is no any ambiguity in the transcription task, only one member of the research group listened to all the audios, segment-to-segment, and provided the corresponding transcriptions.

3.2 Annotation

Once the transcription was obtained, the segments were annotated with an emotional label extracted from the text. In order to obtain a label that only considers textual information, the labelling was carried out by an annotator that was not involved in the transcription process. In this way, the annotator did not listened to the audio previously and was only focused on text. Two members of the research group carried out the annotation independently. The same levels of anger employed in the acoustic annotation was also considered here: zero for very low, one for medium and neutral and two for high degree of anger. Given the ambiguity associated to this annotation procedure a method was also needed to deal with disagreement among the two annotators. Thus, the set of categories consisting of five degrees defined as *very low* agreed by annotators, *low*, which corresponds to a low-medium disagreement, *medium* agreed by annotators, *high*, which corresponds to a medium-high disagreement and *very high* agreed by annotators was also employed here.

An analysis of the labeled segments showed that one of the two annotators always used a higher level of anger than the other one. Therefore, one/zero or two/one labels appeared more frequently than zero/one and one/two labels. It is noteworthy, that segments labeled as zero/two and two/zero (strong disagreement) were very unfrequent. The left side of Table 1 (TEXT-BASED) shows the number of segments for each audio file.

3.3 Analyzing the annoyance

People's behaviour can be very different when regarding the way in which they express annoyance. Some people vary the pitch or intensity of their utterance very easily when they are upset, while others keep these variables unaltered and emphasize the meaning of their message. Thus, both issues have to be considered to identify annoyance rates. For instance, in the audio *Very angry* there is a speech segment where the speaker says: *"sorry, sorry but you are the impolite"*. Regarding the acoustic annotation it was labeled with a low annoyance rate due to an unaltered acoustic signal, whereas the text annotation indicated a high annoyance rate, because its meaning clearly denotes that the user is upset.

Table 1 shows that significant differences can be observed in text-based and speech-based annotations. Mainly, it seems that higher anger rates are associated to speech-based annotations: there are 57 very high segments annotated in speech versus 13 in text, while at the other end, there are 5 very low segments annotated in speech versus 78 in text. An example of this behaviour would be a segment of the audio "Angry 2" where the speaker says: *"all things you will tell me, they have already told me"*. This fragment, was annotated as high in speech annotation and low in textual annotation. It seems, according to the obtained results, that annoyed people tend to vary the features of their utterances (pitch, intensity, etc.) more easily while keeping the meaning of their messages unaltered. It might be because changes in the acoustic signal occur in an spontaneous and involuntary way and it does not happen when regarding the content of the message. Thus, only when the annoyance is kept in a longer period of time annoyance signals appear in text messages. The combination between speech annotator information and text annotator information, provides the chance to complement the information provided from isolated sources, leading to nuanced results, in cases of discrepancy.

A bi-modal annoyance recognizer, combining acoustic and text, was developed in this work and the results provided by its evaluation are given in section 4.

4 Experimental Results

The experiments carried out aimed at analyzing the validity of the assumptions made in the annotation procedure and also the performance when combining acoustic features and textual information. We used the Naive Bayes Classifier (NB) and the Support Vector Machine (SVM) which had already proven their efficiency classifying emotional hints (Irastorza y Torres, 2016). To this end we shuffled the set of frames of each audio file and then split this set into a training and a test set that included 70 % and 30 % of frames, respectively. We used the frame classification accuracy as evaluation metric.

Two series of classification experiments were carried out in order to include textual information in different ways. Firstly, we chose a combination of acoustic (Linear Predic-

	TEXT-BASED					SPEECH-BASED					
	Activation level category										
	v_low	low	medium	high	v_high	v_low	low	medium	high	v_high	Total
Disappointed	1	1	5	2	0	0	6	1	2	0	9
Angry 1	0	4	0	1	1	0	0	4	2	0	6
Angry 2	0	1	2	4	1	1	2	1	4	0	8
Very Angry	65	61	53	35	9	0	95	26	45	57	223
Fed-up	6	2	4	1	2	0	3	11	1	0	15
Impotent	6	3	0	0	0	4	4	1	0	0	9
All	78	72	64	43	13	5	110	44	54	57	270

Table 1: Number of segments for each audio file.

tion Coefficients, LPC) and textual (labels provided in the text annotations) features. The labels obtained from the speech annotations were employed to train the classifiers. In a second stage only LPC acoustic features were considered, but in this case the labels provided to train the classifiers were those obtained from the text annotation. We aimed at analysing the behaviour of the selected sets of features and annotation schemes when classifying frames into the categories that represent the customer annoyance degree in each particular call. Moreover, speaker dependent and speaker independent experiments were also considered. In speaker dependent experiments only frames obtained from a specific speaker (an audio file) were involved in the classification process (training and test).

4.1 First series of experiments

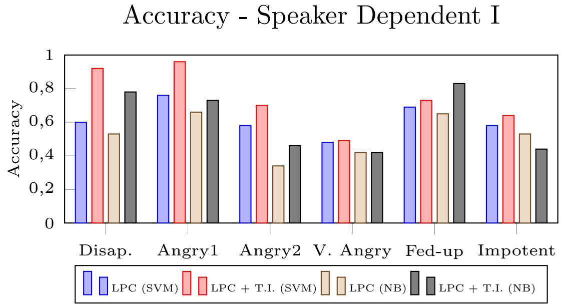

Figure 1: Comparison of SVM and NB frame classification approaches. Bar graphs show the frame identification accuracies based on the sets of features selected for the first series of experiments. The two bar graphs on the left side of each audio correspond to results obtained by SVM classifier whereas the ones on the right side correspond to the results obtained by NB.

In this first set of frame classification experiments, Figure 1 confirms that Linear Prediction Coefficients plus textual information classification outperforms Linear Prediction Coefficients classification in both SVM and NB models for speaker dependent. For instance, we can see an accuracy improvement up to 0.3 in the *Disappointed* audio when combining acoustic parameters plus textual information using SVM classifier. Equally, we carried out speaker independent experiments and the results also showed better performace using acoustic features plus textual information, improving the accuracy from 0.46 in to 0.52. Looking at the results it seems that there is information within text, that is missing in the acoustic signal, that could be useful for detecting annoyance rate.

4.2 Second series of experiments

Then a second set of frame classification experiments was carried out using both SVM and Naive Bayes models. This series was aimed at evaluating speaker dependent/independent model using acoustic parameters with textual labeling.

Figure 2 shows that the classification based on acoustic features (LPC) along with the use of labels based on the text annotation provides lower accuracy values. This loss of accuracy reinforces the idea that cognitive processing diverges depending on which senses are involved in the annotation process. On the other hand, speaker independent results showed also worse performance, since accuracy result did not achieve more than 0.31.

5 *Conclusions and future work*

In conclusion, two main ideas resulted from the experiments. On the one hand, the possibility of joining textual and acoustic information in order to predict annoyance rates was explored and validated. The experiments show that the inclusion of labels extracted from text as a feature improve the classification accuracy. However, using acoustic fea-

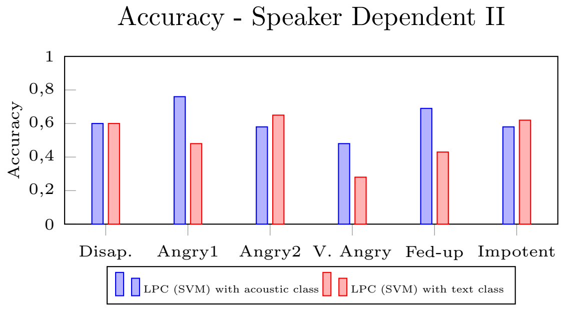

Figure 2: Comparison of SVM and NB frame classification approaches. Bar graphs show the frame identification accuracies based on the sets of features selected for the first second of experiments. The bar graph on the left side of each audio correspond to results obtained by SVM classifier using the class based on the text whereas the right graph correspond to the results obtained using the class based on the speech.

tures with text-based annotation does not provide a good system performance. Acoustic features are linked to acoustic signal, and that is why categories based on semantic analysis are meaningless.

Moreover, the use of a corpus that include spontaneous emotions gathered in a realistic environment leads to an easy extrapolation of the obtained results to a real system.

For further work we propose to explore alternative ways of integrating textual information along with deep learning based classification paradigms.

References

Ben-David, B. M., N. Multani, V. Shakuf, F. Rudzicz, y P. H. H. M. van Lieshout. 2016. Prosody and semantics are separate but not separable channels in the perception of emotional speech: Test for rating of emotions in speech. *Journal of Speech, Language, and Hearing Research*, 59(1):72–89.

Boersma, P. y D. Weenink. 2016. Praat: doing phonetics by computer. Software tool, University of Amsterdam. version 6. 0.15.

Cambria, E. 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107, Mar.

Clavel, C. y Z. Callejas. 2016. Sentiment analysis: From opinion mining to human-agent interaction. *IEEE Transactions on Affective Computing*, 7(1):74–93, Jan.

Devillers, L., L. Vidrascu, y L. Lamel. 2005. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407 – 422. Emotion and Brain.

Eskimez, S. E., K. Imade, N. Yang, M. Sturge-Apple, Z. Duan, y W. Heinzelman. 2016. Emotion classification: how does an automated system compare to naive human coders? En *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, páginas 2274–2278, March.

Esposito, A., A. M. Esposito, L. Likforman-Sulem, M. N. Maldonato, y A. Vinciarelli, 2016. *Recent Advances in Nonlinear Speech Processing*, capítulo On the Significance of Speech Pauses in Depressive Disorders: Results on Read and Spontaneous Narratives, páginas 73–82. Springer International Publishing, Cham.

Eyben, F., M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, y R. Cowie. 2010. On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues. *Journal on Multimodal User Interfaces*, 3(1):7–19, Mar.

Gilbert, E. y K. Karahalios. 2010. Widespread worry and the stock market. En *ICWSM 2010 - Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, páginas 58–65.

Girard, J. M. y J. F. Cohn. 2016. Automated audiovisual depression analysis. *Current Opinion in Psychology*, 4:75 – 79.

Irastorza, J. y M. I. Torres. 2016. Analyzing the expression of annoyance during phone calls to complaint services. En *2016 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, páginas 000103–000106, Oct.

Justo, R., O. Horno, M. Serras, y M. I. Torres. 2014. Tracking emotional hints in spoken interaction. En *Proc. of VIII Jornadas en Tecnología del Habla and IV Iberian SLTech Workshop (IberSpeech 2014)*, páginas 216–226.

Kim, J. C. y M. A. Clements. 2015. Multimodal affect classification at various temporal lengths. *IEEE Transactions on Affective Computing*, 6(4):371–384, Oct.

Marsden, P. V. y K. E. Campbell. 2012. Reflections on conceptualizing and measuring tie strength. *Social Forces*, 91(1):17–23.

Medeiros, L. y C. N. van der Wal. 2017. An agent-based model predicting group emotion and misbehaviours in stranded passengers. En E. Oliveira J. Gama Z. Vale, y H. Lopes Cardoso, editores, *Progress in Artificial Intelligence*, páginas 28–40, Cham. Springer International Publishing.

Meilán, J. J. G., F. Martínez-Sácnhez, J. Carro, D. E. López, L. Millian-Morell, y J. M. Arana. 2014. Speech in alzheimer?s disease: Can temporal and acoustic parameters discriminate dementia? *Dementia and Geriatric Cognitive Disorders*, 37(5-6):327–334.

Mencattini, A., E. Martinelli, F. Ringeval, B. Schuller, y C. D. Natlae. 2016. Continuous estimation of emotions in speech by dynamic cooperative speaker models. *IEEE Transactions on Affective Computing*, PP(99):1–1.

Poria, S., I. Chaturvedi, E. Cambria, y A. Hussain. 2016. Convolutional mkl based multimodal emotion recognition and sentiment analysis. En *2016 IEEE 16th International Conference on Data Mining (ICDM)*, páginas 439–448, Dec.

Ringeval, F., F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, y B. Schuller. 2015. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, 66:22 – 30.

Valstar, M., B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, y M. Pantic. 2014. Avec 2014: 3d dimensional affect and depression recognition challenge. En *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, AVEC '14, páginas 3–10, New York, NY, USA. ACM.

Ververidis, D. y C. Kotropoulos. 2006. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162 – 1181.

Vidrascu, L. y L. Devillers. 2005. Detection of Real-Life Emotions in Call Centers. En *Proceedings of INTERSPEECH'05: the 6th Annual Conference of the International Speech Communication Association*, páginas 1841–1844, Lisbon, Portugal. ISCA.

Wang, K., N. An, B. N. Li, Y. Zhang, y L. Li. 2015. Speech emotion recognition using fourier parameters. *IEEE Transactions on Affective Computing*, 6(1):69–75, Jan.

Wollmer, M., F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, y R. Cowie. 2008. Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. páginas 597–600, 9.