Emotion recognition from the human voice

Conference Paper · April 2013

DOI: 10.1109/SIU.2013.6531196

CITATIONS
2

READS
300

2 authors:

Cevahir Parlak
Bogazici University
4 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Banu Diri
Yildiz Technical University
129 PUBLICATIONS **997** CITATIONS
SEE PROFILE

A Cross-Corpus Experiment in Speech Emotion Recognition

Cevahir Parlak , Banu Diri, Fikret Gürgen

[1] Yildiz Technical University, Turkey
[1] Yildiz Technical University, Turkey
[1] Bogazici University, Turkey
cevahir.parlak@boun.edu.tr, banu@ce.yildiz.edu.tr, gurgen@boun.edu.tr

Abstract

In this work we will introduce EmoSTAR as a new emotional database and perform cross-corpus tests between EmoSTAR and EmoDB (Berlin Emotional Database) using one of the two databases as training set and the other as test set. We will also investigate the performance of feature selectors in both databases. Feature extraction will be implemented with openSMILE toolkit employing Emobase and Emo_large configurations. Classification and feature selection will be run with WEKA tool. EmoSTAR is still under development for more samples and emotion types and we will welcome emotional speech sample donations from the speech community. EmoSTAR is available only for personal research purposes via email to the authors by signing an End User License Agreement.

Index Terms: EmoSTAR, cross-corpus emotion analysis, emotion mining

1. Introduction

We obtained most of the data until today through textual interfaces but the advent of new technologies enabled us to use human speech as a data source. The works on human speech date back to 50 years ago but emotion extraction is relatively a new research field and attracted wide interest. But preparing the necessary data to work on is a very formidable task. Because there are many kinds of emotions and obtaining these speech samples under realistic conditions constitutes very challenging problems. Earlier works in this field dealt only with whether the speech is positive or negative.

Call center applications, computer aided learning systems, lie machines, voiced email systems and voice therapy are some of application areas of emotion extraction Recently in cars, emotion extraction applications are used to detect the relation between the driver's performance and emotional state. Games and human-robot interaction are another application areas for such systems [1]. Emotional speech synthesis is also on its way for a better human-robot interaction [2], [3].

In the second part of this article we mention the related works, in Section 3 EmoSTAR is introduced as a new emotional database, in Section 4 we describe the experiment and show results of single-corpus tests, in Section 5 cross-corpus results are provided, in Section 6 the results of feature selection are given finally in Section 7 we finalize this paper with the conclusions and future works.

2. Related Works

In the last 20 years accelerating works have been done on the emotion extraction. Many emotional databases have been created and on these databases countless applications have been developed [4], [5], [6]. At the beginning of these works only positive and negative classification has been implemented. But as technology matured more subtle classifications were tried on increasing numbers of emotion types and very high levels of recognition rates have been reached. But one of the major problems we face today in these applications is the low recognition rates of cross-corpus implementations. Cross-corpus, blend-corpus and multi-corpus works [7-17] are relatively new and scarce compared to the single-corpus works.

Oflazoglu and Yildirim [8] used TURES (Turkish Emotional Speech Database) and VAM (Vera Am Mittag) for cross-corpus experiments and reported 41.3% success rate in three dimensional emotion space.

In [9] Zhang et al. found 62.6% for arousal and 55.6% for valence in their cross-corpus experiments on ABC, AVIC, DES, eNTERFACE, SAL and VAM databases.

In [11] Schuller et al. works on AVIC, DES, EmoDB, eNTERFACE, SmartKOM, SUSAS yielded 35% to 45% success rates for EmoDB as test set and others as training set.

In [12] 8 databases ABC, AVIC, DES, EmoDB, eNTERFACE, SAL, SUSAS and VAM used by Schuller et al. for cross-corpus evaluation . They used one database as training set and all the other 7 datasets as test sets. Tarining on VAM gives 67.7% mean unweighted accuracy in arousal and training on DES gives 54.8% mean unweighted accuracy in valence states of dimensional classifications.

In [13] Eyben et al. reported 53.4% mean unweighted accuracy with SVM (SMO) over SmartKOM, Aibo, SAL and VAM corpora using three of the corpora as training set and the remaining corpus as test set.

3. EmoSTAR

In the recent years many works have been conducted on emotion extraction and many emotional databases have been created. In this work we created our own dataset from TV and internet recordings. Although obtaining neutral samples from TV is easy and very abundant (e.g. news channels), it is very difficult to obtain other emotional speech samples. Even sometimes music or other noise accompanies the scene and worsens the case further. On the other hand it is difficult to direct the participants to speak naturally when we recruit them to record speech by acting.

First of our speech database is EmoSTAR which consists of 393 angry, neutral, happy and sad speech utterances in Turkish and English language. Neutral utterances have been recorded mainly from news channels, angry utterances collected from TV movies, sad utterances obtained from internet videos and happy utterances have been collected from the speeches of award winners in various ceremonies such as Oscar, Golden Globe and American Music Awards. Happy and neutral utterances are natural, angry utterances are acted and sad utterances are also natural except a few acted samples.

All utterances contain different sentences and lengths of speech samples vary between 2.2 and 14.5 seconds. Number of sentences by speaker is varying from 1 to 29. Categorical annotation was conducted by the author with audio-visual evaluation in accordance with the classification of the owner of the videos who labeled them as sad or angry. Detailed numbers are shown in Table 1. EmoSTAR is still under development to contain more samples and emotion types and available via email to the authors by signing an End User License Agreement.

Table 1. Emotions in EmoSTAR (E=English, T=Turkish).

	Angry	Neutral	Happy	Sad
Male	33 E 30 T	35 E 34 T	45 E	12 E
Female	40 E	37 E 20 T	37 E	51 E 19 T
Total=393	103	126	82	82

The next database on which we will work is Berlin Emotional Database (EmoDB) which consists of 535 speech utterances of 127 angry, 79 neutral, 71 happy, 62 sad, 69 fear, 81 bored and 46 disgust as shown in Table 2. These utterances are in German language and there are 5 short and 5 long sentences. Pan [5] achieved 95.1% success rate for happy, neutral and sad discrimination in EmoDB. The same works were conducted on five emotion types of happy, sad, neutral, bored and disgust. Energy and Prosody features achieved 66%, LPCMCC (Linear Prediction Coding Mel Cepstrum Coefficients) 70%, both combined achieved 82% success rate. Wu et al. [6] achieved 85.6% weighted average recall rates on EmoDB in classification of 7 emotions using SVM classifier with 10-fold cross-validation.

Table 2. Emotions in EmoDB.

	A	N	H	S	B	D	F
Male	60	39	27	25	35	11	36
Female	67	40	44	37	46	35	33
Total=535	127	79	71	62	81	46	69

4. Experimental Setup

Experiments will be conducted with openSMILE [18] and Weka [19] tools. Feature extraction will be run using Emobase and Emo_large configuration files of openSMILE. Number of features in these file are shown in Table 3.

Table 3. Feature numbers in openSMILE configuration files.

Name	Feature Numbers
Emobase.conf	988 (26 LLD + 26 delta)*19 functionals
Emo_large.conf	6669 (57 LLD + 57 delta+ 57 delta-delta)*39 functionals

All experiments will be run with 10-fold cross-validation using Naive Bayes (NB), SMO and Bagging (Bag) classifiers of Weka tool. Weighted average classification results in EmoDB and EmoSTAR are shown in Table 4 (7 emotions) and Table 5 (4 emotions: angry, neutral, happy, sad) and Table 6 using Emobase and Emo_large configurations.

Table 4. Classification results for 7 emotions in EmoDB.

EmoDB	Emobase (988)	Emo_large (6669)
NB	53.45	70.09
SMO	**87.28**	86.54
Bag	67.28	71.40

Table 5. Classification results for 4 emotions (Angry, Neutral, Happy, Sad) in EmoDB.

EmoDB	Emobase (988)	Emo_large (6669)
NB	80.82	84.66
SMO	**92.33**	90.26
Bag	81.41	80.53

Table 6. Classification results for 4 emotions (Angry, Neutral, Happy, Sad) in EmoSTAR.

EmoSTAR	Emobase (988)	Emo_large (6669)
NB	83.71	83.96
SMO	**96.18**	**96.94**
Bag	84.47	87.78

In the next step we mix EmoSTAR and EmoDB and monitor the performances shown in Table 7.

Table 7. Classification results for 4 emotions (Angry, Neutral, Happy, Sad) in EmoSTAR+EmoDB.

EmoSTAR+EmoDB	Emobase (988)	Emo_large (6669)
NB	63.52	79.64
SMO	90.71	**90.98**
Bag	81.69	82.92

5. Cross-Corpus Experiments

After 20 years of works on emotion extraction we are very close to the limits of accuracy in single-corpus experiments. Cross-corpus experiments pose another challenge for the researchers to overcome regarding sometimes very low success rates in these applications. We run cross-corpus experiments using one database as training set and the other as test set. Since EmoSTAR has 4 emotion types, cross-corpus implementations are executed by using 4 emotions of EmoDB. Different groupings of these 4 emotions are also examined such as angry-neutral and neutral-happy-sad as in Table 8 and Table 9.

Table 8. EmoSTAR and EmoDB as training and test sets for different emotion groupings in Emobase configuration.

	EmoSTAR Training EmoDB Test Emobase (988)			EmoDB Training EmoSTAR Test Emobase (988)		
	NB	**SMO**	**Bag**	**NB**	**SMO**	**Bag**
A,N	55.82	**74.75**	74.27	51.09	65.50	55.89
A,N,H,S	41.88	**47.19**	40.41	40.96	44.27	35.36
N,H,S	51.58	51.41	**62.73**	39.65	54.13	53.44

A: Angry, N: Neutral, H: Happy, S: Sad

Table 9. EmoSTAR and EmoDB as training and test sets for different emotion groupings in Emo_large configuration.

	EmoSTAR Training EmoDB Test Emo_large (6669)			**EmoDB Training EmoSTAR Test Emo_large (6669)**		
	NB	**SMO**	**Bag**	**NB**	**SMO**	**Bag**
A,N	71.84	**91.26**	75.24	53.27	72.05	49.78
A,N,H,S	45.13	**62.83**	34.51	42.74	44.02	32.82
N,H,S	45.75	52.83	**53.77**	48.96	52.75	33.44

6. Feature Selection

It is the holy grail of researchers to find the optimal set of features in the classification tasks. It has been reported that redundant features may hinder the performance of classification algorithms. This idea is proven in this work so far as seen from Table 4, 5 and 6. In Table 5 the performance of SMO which is the best classifier dropped more than 2%. Feature selection algorithms are ready to address this issue. Feature selection has been implemented via Weka's Command Line Interface (CLI) using the following command:

java weka.filters.unsupervised.attribute.Remove -V -R 1-4, 29, 45 -i emodb1.arff -o emodb2.arff

This command tells Weka to remove all features except feature 1, 2, 3, 4, 29 and 45 from emodb1.arff and write the output file to emodb2.arff. The associated numbers of features are extracted by running attribute selectors in "Use full training set" option. We can avoid manually selecting such huge numbers of features by simply copying and pasting these numbers to the Weka CLI. We will employ 4 types of feature selectors.

- Information Gain Attribute Evaluator + Ranker,
- CfsSubSet Evaluator + Linear Forward Selection,
- ChiSquared Attribute Evaluator + Ranker,
- Principal Component Analysis + Ranker

search methods are chosen as candidate feature selectors and results are presented in Table 10 and Table 11 for EmoDB and in Table 12 and Table 13 for EmoSTAR.

Table 10. Feature selection for EmoDB (Emobase).

EmoDB	**Emobase (988)** ‡		
	NB (53.45)*	**SMO** (87.28)*	**Bag** (67.28)*
CfsSub+LFS (56) †	75.88	80.37	69.71
InfoG+Rank (508) †	75.88	87.10	69.34
Chi+Rank (716) †	75.14	**88.03**	85.24
PCA (180) †	66.16	80.74	65.60

* Results with the full feature set
† Numbers of selected features
‡ Numbers of full features

In EmoDB, Chi+Rank achieved 88.03% success rate using only 716 features out of 988. This accuracy is above the success rate of the full feature set with 988 features and also above the Emo_large set with huge 6669 features. Chi+Rank achieved same success in Emo_large configuration and surpassed the full feature set. All feature selectors provided a big surge of 22% in Naive Bayes classifier when Emobase configuration is utilized in EmoDB as seen from Table 10. This is a clear indication of degradation of performance with redundant features.

Table 11. Feature selection for EmoDB (Emo_large).

EmoDB	**Emo_large** (6669) ‡		
	NB (70.09)*	**SMO** (86.54)*	**Bag** (71.40)*
CfsSub+LFS (94) †	77.0	82.99	71.40
InfoG+Rank (4407) †	71.77	87.28	74.01
Chi+Rank (5417) †	71.21	**87.85**	72.71

In EmoSTAR, InfoG+Rank and Chi+Rank surpassed all others with 97.45% accuracy as shown in Table 13. Feature selectors surpassed the full feature set in terms of accuracy in all classifications.

PCA demonstrated very poor performances compared to the others and we were unable to implement it for Emo_large configurations. Even by selecting 180 features, PCA is below the CfsSubset Evaluator which uses the least number of features among the 4 feature selectors as indicated in Table 10 and Table 12. But we should keep in mind that PCA implementation has a very serious drawback. It creates many feature groups containing different combinations as output and selecting the one that yields the highest score is very difficult.

Table 12. Feature selection for EmoSTAR (Emobase).

EmoSTAR	**Emobase** (988) ‡		
	NB (83.71)*	**SMO** (96.18)*	**Bag** (84.47)*
CfsSub+LFS (67) †	87.27	94.40	86.25
InfoG+Rank (837) †	83.96	**96.69**	85.24
Chi+Rank (837) †	83.96	**96.69**	85.24
PCA (180) †	75.06	89.56	79.64

Table 13. Feature selection for EmoSTAR (Emo_large).

EmoSTAR	**Emo_large** (6669) ‡		
	NB (83.96)*	**SMO** (96.94)*	**Bag** (87.78)*
CfsSub+LFS (87) †	87.78	94.14	88.04
InfoG+Rank (5637) †	83.20	**97.45**	88.80
Chi+Rank (5637) †	82.69	**97.45**	88.80

7. Conclusions

One of the major conclusions of this paper is the moderate recognition rates of cross-corpus results compared to the single-corpus and blend-corpus results.

The Emo_large feature set which employs 6669 features demonstrated stronger performances especially in cross-corpus experiments when EmoSTAR is used as training set. But in

single-corpus tests with EmoDB it is below the performance of the Emobase set.

Another promising conclusion is the extremely good performances of feature selectors. Feature selectors are able to surpass the original full feature set regardless of the enormous numbers of features. In EmoDB and EmoSTAR, we are able to surpass the success rate of Emo_large configuration using nearly 1/10th of the features of the original full feature set.

In the future works we will develop EmoSTAR for more emotional speech samples and more emotion types and we plan to conduct the experiments with more databases. We will kindly accept emotional speech sample contributions from the speech community to enlarge the database.

8. References

[1] Ramakrishnan, S., "Recognition of Emotion from Speech: A Review", International Journal of Speech Technology, v:15, Issue 2, pp 99-117, 2012.

[2] Iida, A., Campbell, N., Higuchi, F. and Yasumura M., "A corpus-based speech synthesis system with emotion", Speech Communication 40 161–187, 2003.

[3] Black, A. W., Bunnel, H.T., Dou, Y., Kumar, P., Metze, F., Perry, D., Polzehl, T., Prahallad, K., Steidl, S. and Vaughn, C., "New Parameterization for Emotional Speech Synthesis", CSLP Proc., Johns Hopkins Summer Workshop, Baltimore, 2011.

[4] Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G. and Wendemuth, A., "Acoustic Emotion Recognition: A Benchmark Comparison of Performances", IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2009 Proc., Merano, Dec. 2009.

[5] Pan, Y., Shen, P. and Shen, L., "Speech Emotion Recognition Using Support Vector Machine", International Journal of Smart Home, v:6, no. 2, Apr. 2012.

[6] Wu, S., Falk T.H. and Chan W., (2010). "Automatic speech emotion recognition using modulation spectral features", Speech Communication 2010, doi:10.1016/j.specom.2010.08.013.

[7] Bone, D., Lee, C. and Narayanan, S.S., "A Robust Unsupervised Arousal Rating Framework using Prosody with Cross-Corpora Evaluation", INTERSPEECH 2012 Proc., Oregon, Sept. 2012.

[8] Oflazoglu C. and Yildirim S., "Recognizing emotion from Turkish speech using acoustic features", EURASIP Journal on Audio, Speech, and Music Processing, 2013.

[9] Zhang, Z., Weninger, F., Wöllmer, M. and Schuller, B., "Unsupervised Learning in Cross-Corpus Acoustic Emotion Recognition", IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU Proc., Waikoloa, Hawaii, Dec. 2011.

[10] Shami, M. and Verhelst, W., "Automatic classification of expressiveness in speech: a multi-corpus study. in Speaker Classification", II LNCS Proc., ed. by Müller C (Berlin: Springer), pp. 43–56, 2007.

[11] Schuller, B., Vlasenko, B., Eyben, F., Wollmer, M., Stuhlsatz, A.,Wendemuth, A. and Rigoll, G., "Cross-corpus acoustic emotion recognition: variances and strategies", IEEE Transactions on Affective Computing 1(2), 119–131, July-December 2010.

[12] Schuller, B., Zhang, Z., Weninger, F. and Rigoll, G., "Selecting Training Data for Cross-Corpus Speech Emotion Recognition: Prototypicality vs. Generalization", 2011 Speech Processing Conference, AVIOS Proc., Telaviv, June 2011.

[13] Eyben, F., Batliner, A., Schuller, B., Seppi, D. and Steidl, S., "Cross-Corpus Classification of Realistic Emotions – Some Pilot Experiments", 7th Intern. Conf. on Language Resources and Evaluation LREC Proc. 2010, Valletta, ELRA, 2010. 19.-21.05.2010.

[14] Marchi, E., Batliner, A., Schuller, B., Fridenzon, S., Tal, S. and Golan, O., "Speech, Emotion, Age, Language, Task, and Typicality: Trying to Disentangle Performance and Feature Relevance", ASE/IEEE International Conference on Social Computing and ASE/IEEE International Conference on Privacy, Security, Risk and Trust Proc., Maryland, Apr. 2012.

[15] Schuller, B., Zhang, Z., Weninger, F. and Rigoll, G., "Using Multiple Databases for Training in Emotion Recognition: To Unite or to Vote?", INTERSPEECH 2011 Proc., Florence, August 2011.

[16] Tahon, M., Delaborde, A. and Devillers, L., "Real-life Emotion Detection from Speech in Human-Robot Interaction: Experiments Across Diverse Corpora with Child and Adult Voices", ISCA, INTERSPEECH 2011 Proc., Florence, August 2011.

[17] Shami, M. and Verhelst, W., "An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech," Speech Communication, 2007.

[18] Eyben, F., Wöllmer, M. ve Schuller, S., (2009). "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor", In Proc. ACM Multimedia (MM), ACM, Florence, Italy, ACM, ISBN 978-1-60558-933-6, pp. 1459-1462, October 2010. doi:10.1145/1873951.1874246

[19] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., (2009). "The WEKA data mining software: an update". SIGKDD Explor. Newsl. 11, 10–18 (2009).