Investigating the Corpus Independence of the Bag-of-Audio-Words Approach

Mercedes Vetráb[1,2](✉) and Gábor Gosztolya[1,2]

[1] Institute of Informatics, University of Szeged, Árpád tér 2, Szeged, Hungary
{vetrabm,ggabor}@inf.u-szeged.hu

[2] MTA-SZTE Research Group on Artificial Intelligence, Tisza Lajos körút 103, Szeged, Hungary

Abstract. In this paper, we analyze the general use of the Bag-of-Audio-Words (BoAW) feature extraction method. This technique allows us to handle the problem of varying length recordings. The first step of the BoAW method is to define cluster centers (called codewords) over our feature set with an unsupervised training method (such as k-means clustering or even random sampling). This step is normally performed on the training set of the actual database, but this approach has its own drawbacks: we have to create new codewords for each data set and this increases the computing time and it can lead to over-fitting. Here, we analyse how much the codebook depends on the given corpus. In our experiments, we work with three databases: a Hungarian emotion database, a German emotion database and a general Hungarian speech database. We experiment with constructing a set of codewords on each of these databases, and examine how the classification accuracy scores vary on the Hungarian emotion database. According to our results, the classification performance was similar in each case, which suggests that the Bag-of-Audio-Words codebook is practically corpus-independent. This corpus-independence allows us to reuse codebooks created on different datasets, which can make it easier to use the BoAW method in practice.

Keywords: Emotion detection · Bag-of-Audio-words · Human voice · Sound processing

1 Introduction

Human speech is not only used for encoding the words uttered, but it also includes some information about the speakers physical and mental state. One of the latter attributes is the emotional state of the speaker. Nowadays emotion detection from audio data (speech emotion recognition, SER) is an active area of research with a wide range of possible applications, including human-computer interfaces (monitoring human communication) [6], dialog systems [1] and call centers [12]. In the future with good emotion recognition systems, we will be able to create more human-oriented and friendlier systems.

P. Sojka et al. (Eds.): TSD 2020, LNAI 12284, pp. 285–293, 2020.
https://doi.org/10.1007/978-3-030-58323-1_31

Since the beginning of research in this area, many feature extraction and classification techniques have been used along with different datasets to get the best results. The basis of our study is a previous paper [11], where we investigated the Bag-of-Audio-Words (BoAW [7]) technique and its efficiency. One of the major problem using the BoAW technique was the time required to generate a codebook, which could be solved if we utilize a predefined codebook instead of generating a new one for each data set. In this paper, we discuss the consequences of using a predefined codebook. We address the question of whether a codebook from another database can produce similar or better results than by using a codebook from the original database. We perform our experiments on a Hungarian emotion speech database; previous classification accuracy scores on this database were around 66–70%. We measured Unweighted Average Recall (UAR, [9]) scores in the range 66–71%, so our view is that the BoAW method with a predefined codebook is a competitive technique for emotion recognition.

2 The Bag-of-Audio-Words Method

With the representation of emotional speech data, there are many open questions and problems. One of them is feature extraction from recordings. Often the utterances we have to handle are of different lengths, but most classification techniques require fixed-sized feature vectors. The Bag-of-Audio-words is a feature extraction method similar to the Bag-of-Words [7] technique. With the BoAW feature representation, we can resolve the problem of varying length.

In the BoAW procedure, first we have to extract the frame-level feature vectors per recording; unfortunately, the number of vectors created depends on the original length of the evaluated recording and the frame's windowing size. In the next step, we collect all the feature vectors from all the recordings of the training set, put them into one big "bag" and perform clustering on it. Cluster size (N) is one of the parameters of the BoAW method. The result of the clustering step, the center vector of each calculated cluster, is called a "codeword". The group of codewords is then called the "codebook".

After, in the vector quantization step, we again work with individual recordings and create a histogram for each recording (both for the training and test sets). We calculate the closest codeword for each feature vector in the actual recording and replace the original feature vectors by the index of the closest codeword. We can also specify how many closest vectors we examine (this is also a parameter of the BoAW method). As a result, the same sized (i.e. N) histogram is produced for each recording. All of the codeword indices appear on the histogram's x axis. On the y axis, there are quantities which represent the set of recording feature vectors that were mapped to a particular codeword.

In the last step, we normalize the histogram, so the given frequencies are divided by the number of frames of the speech recording. These normalized histograms will be our new feature vectors, that have an independent length from the recording sizes (i.e. they will consist of N values) We will call this set of histograms "Bag-of-Audio-Words" and use it as features for our classifier.

3 Data and Methods

3.1 Data Sets

In each experiment, we created and evaluated our classification model on the Hungarian emotion database training and test sets. The other two databases were used to construct the codebook.

Hungarian Emotion Database. This database contains speech from 97 native Hungarian speakers [10]. Most of the segments were recorded from a continuous, spontaneous speaking television program with actors, while the other part came from an improvisation show. In the first case, the samples are vivid, and the emotions are more clear because of the actors. The samples from the second case, however, are closer to real-life emotions. The database contains 1111 sentences, separated into an 831 sample training set (cca. 20 min long) and a 280 sample test set (cca. 7 min long). We had four emotions: neutral, joy, anger and sad.

German Emotion Database. This database (also known as *EmoDB*) contains speech from 10 native German speakers [2]. The recordings were made with actors aged between 25 and 35. Each participant produced 10 German utterances (5 short and 5 longer sentences), all of them with a different emotion. The classification labels were: neutral, anger, boredom, disgust, fear, happiness and sadness. The whole database contains approximately 25 min of recordings.

Hungarian Speech Database. This database contains Hungarian television news recordings taken from 8 different TV channels [5]. The whole data set consists of 28 h of recordings. In terms of emotion detection, all of the labels can be treated as neutral because newsreaders are not allowed to show any emotion.

3.2 Feature Set

Our frame-level feature set is based on the Interspeech ComParE Challenge [9]. This set contains 65 frame-level features (4 energy-related, 55 spectral and 6 voicing related). We used the open-source openSMILE feature extractor [4] with the `IS13 ComParE` config file. For each frame we calculated the derivatives (i.e. Δ values) as well; these hold information about the dynamics of the samples.

3.3 Evaluation

Classification is performed by the LIBSVM library [3]. We optimized the SVM C complexity parameter in the range 10^{-5}, 10^{-4} to 10^{0}. We applied standardization on the BoAW feature vectors before each model was trained. In the optimization part of our experiments, we worked with the training set, based on speaker-independent 10-fold cross-validation. In the test scenario, we trained one SVM model on the whole training set with the optimal C parameter found above and evaluated it on the test set.

Table 1. *Baseline*: best results got with normalization and standardization, when we evaluate our technique with cross-validation and do it on the test set.

Feature-transformation		UAR		Codebook size
	a	CV	Test	
Normalization	5	58.08%	48.13%	512
	10	57.48%	50.27%	512
Standardization	5	55.43%	53.54%	512
	10	56.57%	64.32%	256

3.4 Parameters of the BoAW Method

The BoAW method has many adjustable parameters. In our study, we tested the effect of the preprocessing method, the codebook size N, and the quantization neighbour number parameters on the learning algorithm performance. For the codebook building we used an open-source program called openXBOW [8].

Codebook size: In each experiment we tested the effect of the following lengths: 32, 64, 128, 256, 512, 1 024, 2 048.

Histogram neighbour number: Instead of looking for just the closest codeword, each vector may also be assigned to a certain number of the closest codewords. Previously [11] we found that using more neighbours leads to a more precise description of the recordings besides the same feature vector size. This is why we experimented with two different settings (5 and 10).

Preprocessing techniques: If some of the features have an extremely high or low value compared to the others, it may dominate the Euclidean distance during the BoAW vector quantization step. Previously [11] we found that preprocessing the frame-level vectors by standardizing or normalizing them can improve the performance, so we tried both solutions.

Derivatives: In a previous study [11] we found that using the derivatives of the frame-level attributes can improve the performance, so we also used them in our experiments. The openXBOW tool also gives the opportunity to create separate codebooks for the original frame-level values and another for the Δs; because we opted for this technique, the codebook sizes provided have to be multiplied by 2 to get the actual number of features.

4 Tests and Results

As the baseline, we create the codebook from the Hungarian emotion database training set. Our results are shown on Table 1. The best result of cross-validation (i.e. 58.08%) came with normalization, 5 neighbours, and a 512-sized codebook. The best result of the test (i.e. 64.32%) came with standardization, 10 neighbours and $N = 512$. In addition, it is clear that in 3 out of 4 cases the results obtained on the test database were lower than the results of cross-validation, which may be due to overfitting to the training set during codebook creation.

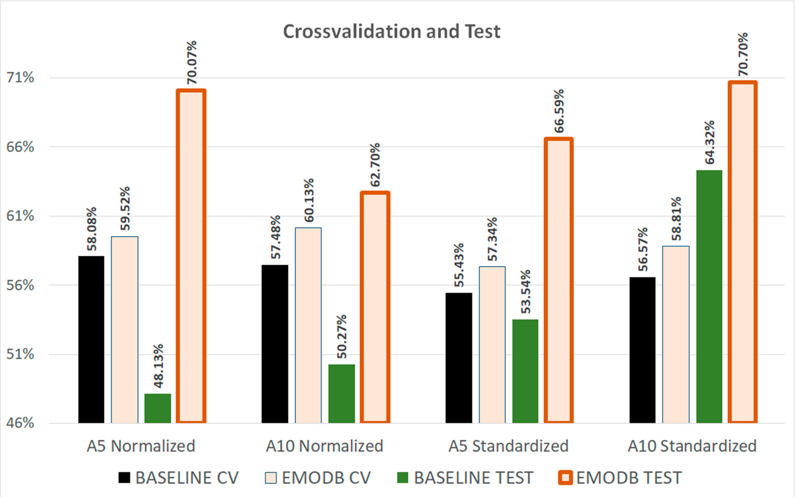

Fig. 1. Results of the *Baseline* and *EmoDB* generated codebooks with cross-validation and evaluation on the test set.

Table 2. *EMODB*: best results with normalization and standardization, when we evaluate our technique with cross-validation and do it on the test set.

Feature-transformation		UAR		Codebook size
	a	CV	Test	
Normalization	5	59.52%	70.07%	1 024
	10	60.13%	62.70%	256
Standardization	5	57.34%	66.59%	128
	10	58.81%	70.70%	256

4.1 Codebook from the *EmoDB* Database

Next, we wanted to know whether working with a codebook from other databases could produce similar or better results than a codebook created from the original database. In this part, the codebooks were created from *EmoDB*; then we built the BoAW representation for the Hungarian emotion database and performed classification using these features.

Examining the results (see Fig. 1 and Table 2) we can see that there was a significant improvement over the baseline in all four test cases. In 2 cases out of 4, we also see a reduction in the size of the required codebook, which can also reduce the time needed to produce a BoAW representation.

This improvement and the fact that all *EmoDB* test cases have more accurate scores than all the *EmoDB* cross-validation scores, in our opinion, might indicate that a codebook made from the original database tends to lead to overfitting,

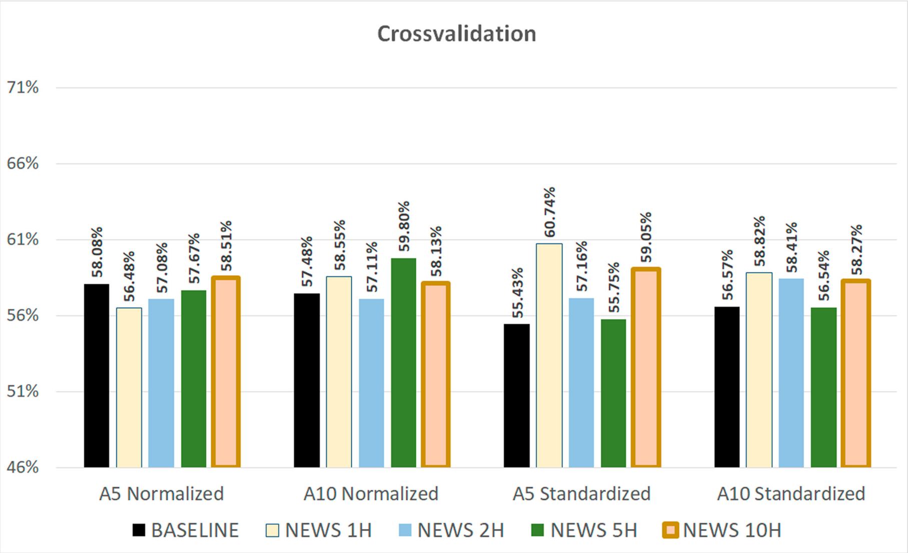

Fig. 2. Cross-validation results got from the Baseline and from the *News* database

and a predefined codebook (which is generated independently from the actual training samples) can eliminate this problem.

4.2 Codebook from the *News* Database

Based on the previous tests, it is apparent that a codebook created from a different database led to significant improvements. On the other hand, it is still not clear whether the type of speech (e.g. rich of emotions or completely neutral) present in the database used for codebook creation affects the emotion classification performance. To examine this, next the codebooks were prepared from subsets of the (non-emotion) Hungarian television recordings database [5]. Otherwise, all classification steps were done similarly as before. To investigate whether the length of the database also affects this performance, we used an 1-h, 2-h, 5-h, and 10-h long subset for codebook creation.

Based on the results of the cross-validation (see Fig. 2), we could not correlate the length of the database with the success of the classification. The same can be said about the type of preprocessing method and the number of closest neighbors: all the scores ranged from 55.75% to 60.74%. Most of the best-performing codebook sizes were 1 024 and 2 048, which are relatively large feature sets. The best result of cross-validation came from a 1-h length database, with standardization, taking 5 neighbors, using 1 024 sized codebook, giving the score of 60.74%. The results did not reveal significant differences depending on the length of the database, hence no general relationship could be found. In addition, we did not get significantly better or worse scores than using a codebook specifically designed for emotion detection from a *EmoDB* database.

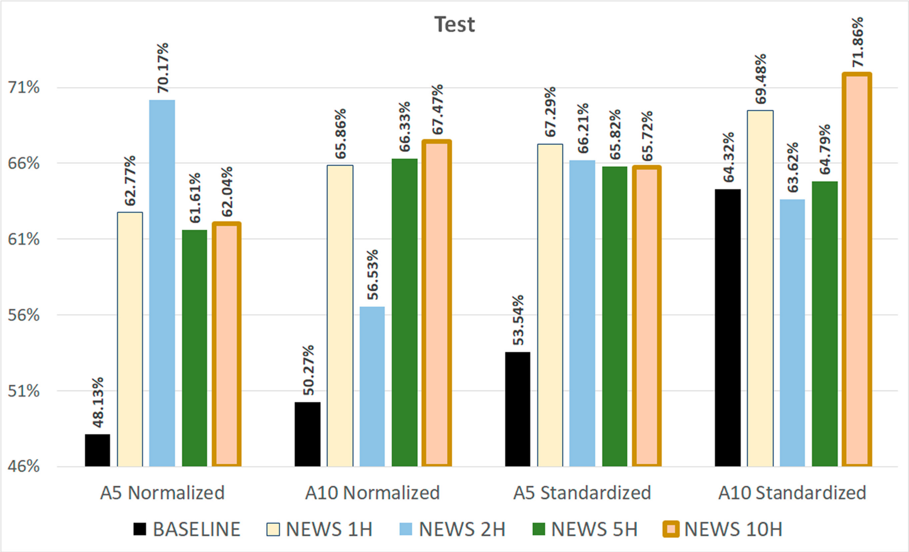

Fig. 3. Test result got from the Baseline and from the Hungarian speech database

The best score of the test was 71.86%, with the 10-h dataset, standardization, with 10 neighbours and a 1 024-sized codebook. However, besides the required increase in the codebook size, no obvious inferences could be made here (Fig. 3).

5 Conclusions

In this paper, the BoAW *(Bag-of-Audio-Words)* feature representation method was simultaneously applied on multiple databases for emotion recognition. We were interested in the possibility of creating BoAW codebooks from other datasets; this would allow the re-using of codebooks for several corpora, therefore allowing to cut execution times significantly. From this viewpoint, building a codebook from other, similar purpose databases gives better scores than those got using purpose-built database codebooks.

Based on our tests, it can be clearly stated that each predefined codebook can be successfully used to extract BoAW feature representations of any other databases. The best score of the tests with the Hungarian emotion database own codebook was 64.32%. Compared to this, when we used other database codebooks we got better results. The best score of the tests with the Hungarian speech database codebook was 66–71.86%. The best score of the tests with the German emotion database codebook was 66–70.70%. With these results, we could not find a clear answer to whether it is advisable to use a codebook between any two databases created for similar purposes but a different language or for a similar language but different purpose. In both cases, our results varied on a similar scale, with no significant difference. They just differed in codebook size, so this point requires deeper study.

Now we know that codebooks are portable, but there are several directions we can pursue in the future. One good question is what type of databases can be most effectively transferred from the viewpoint of codebook reusability. Is there a close connection between certain types of databases? We could also test other frame-level feature sets to see whether there are any benefits in practice.

Acknowledgements. This study was supported by the project "Integrated program for training a new generation of scientists in the field of computer science", grant no. EFOP-3.6.3-VEKOP-16-2017-0002. This study was also supported by the National Research, Development and Innovation Office of Hungary via contract NKFIH FK-124413. This research was also supported by the grant TUDFO/47138-1/2019-ITM of the Hungarian Ministry for Innovation and Technology. Gábor Gosztolya was also funded by the János Bolyai Scholarship of the Hungarian Academy of Sciences and by the Hungarian Ministry of Innovation and Technology New National Excellence Program ÚNKP-19-4-SZTE-51.

References

1. Burkhardt, F., van Ballegooy, M., Engelbrecht, K.P., Polzehl, T., Stegmann, J.: Emotion detection in dialog systems: applications, strategies and challenges. In: Proceedings of ACII, Amsterdam, Netherlands, pp. 985–989 (2009)
2. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A database of German emotional speech. In: Proceedings of Interspeech, pp. 1517–1520 (2005)
3. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 1–27 (2011)
4. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the Munich versatile and fast open-source audio feature extractor. In: Proceedings of ACM Multimedia, New York, NY, USA, pp. 1459–1462 (2010)
5. Tóth, L., Grósz, T.: A comparison of deep neural network training methods for large vocabulary speech recognition. In: Habernal, I., Matoušek, V. (eds.) TSD 2013. LNCS (LNAI), vol. 8082, pp. 36–43. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40585-3_6
6. James, J., Tian, L., Inez Watson, C.: An open source emotional speech corpus for human robot interaction applications. In: Proceedings of Interspeech, Hyderabad, India, pp. 2768–2772 (2018)
7. Pancoast, S., Akbacak, M.: Bag-of-Audio-Words approach for multimedia event classification. In: Proceedings of Interspeech, Portland, USA, pp. 2105–2108 (2012)
8. Schmitt, M., Schuller, B.: openXBOW - Introducing the Passau open-source cross-modal Bag-of-Words toolkit. J. Mach. Learn. Res. **18**(96), 1–5 (2017). http://jmlr.org/papers/v18/17-113.html
9. Schuller, B., et al.: The Interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In: Proceedings of Interspeech, pp. 148–152 (2013)

10. Sztahó, D., Imre, V., Vicsi, K.: Automatic classification of emotions in spontaneous speech. In: Esposito, A., Vinciarelli, A., Vicsi, K., Pelachaud, C., Nijholt, A. (eds.) Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues. LNCS, vol. 6800, pp. 229–239. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-25775-9_23
11. Vetráb, M., Gosztolya, G.: érzelmek felismerése magyar nyelvű hangfelvételekből akusztikus szózsák jellemzőreprezentáció alkalmazásával. In: Proceedings of MSZNY, Szeged, Hungary, pp. 265–274 (2019)
12. Vidrascu, L., Devillers, L.: Detection of real-life emotions in call centers. In: Proceedings of Interspeech, Lisbon, Portugal, pp. 1841–1844 (2005)