THE RELATION OF HUBS TO THE DODDINGTON ZOO IN SPEAKER VERIFICATION

Dominik Schnitzer, Arthur Flexer, Jan Schlüter

Austrian Research Institute for Artificial Intelligence (OFAI)
Freyung 6/6, 1010 Vienna, Austria

ABSTRACT

In speaker verification systems there exists the well-known phenomenon of speakers which are very problematic to verify and have been given various metaphoric animal names. Our work connects this so-called 'Doddington zoo' and the animals of the whole 'biometric menagerie' to the problem of 'hubs' in high dimensional data spaces, which was recently the topic of a number of publications in the machine learning literature. Due to a general problem of measuring distances in high dimensional data spaces, hub objects emerge which have a high similarity to a large number of data items. This is a novel aspect of the 'curse of dimensionality' which adversely affects classification and identification performance. In a series of experiments we try to understand the 'Doddington zoo' problem with respect to the notions of hubs and anti-hubs.

***Index Terms*—** Speaker verification, normalization, hubs

1. INTRODUCTION

In a recent publication by Radovanović et al. [1] the so-called 'hubness' phenomenon has been described and explored as a general problem of machine learning in high dimensional data spaces. 'Hubs' are data points which keep appearing unwontedly often in nearest neighbor lists of many other data points due to a general problem of measuring distances in high dimensions. This effect is particularly problematic for algorithms computing similarity, as the same few objects are found to be similar to a large percentage of objects in a database, and, at the same time, other objects vanish from all neighborhoods.

In speaker-verification systems, the distances between statistical models of spoken words from individual speakers and samples from the same 'genuine' speakers as well as from other 'impostor' speakers are being computed. These similarity scores are used for performance evaluation of the overall system. In the case of speaker-verification systems the 'hubness' problem would manifest itself in the form of impostor audio samples which are similar to a large number of speaker-models, impairing the performance of the system significantly. This effect is, in the field of speaker verification algorithms and more broadly in biometric identification systems, actually known since the nineties and usually referenced as the 'Doddington zoo' effect. Doddington et al. [2] coined the term because they divided the speakers in a verification system into 'wolves', 'sheep', 'lambs' and 'goats'. Audio samples from 'wolves' easily impersonate other speakers (have a high similarity to a lot of speaker models), while persons which are difficult to recognize are denoted as 'goats'. Later on the zoo was even extended with the 'biometric menagerie' [3], adding four more animals. The appearance of such objects was always seen as an isolated problem of verification systems [3] with the cause of this problem unclear as, for example, Poh and Tisarelli [4] note in a recent publication.

This paper places the emergence of Doddington's zoo in a more general context, the phenomenon of hubs, and tries to explain the mechanisms behind them by conducting a series of experiments on speaker verification.

2. RELATED WORK

Objects, being unwontedly often similar to a large number of objects or never similar to any object, have already been observed in many areas besides verification systems. But unfortunately these problems were always seen as an isolated issue of the data or algorithm used. For example, the problem is referenced in image retrieval [5] as 'too-often-selected' and 'never-seen'-images, in music information retrieval [6] ('hub'/'orphan' music pieces) and, most recently, in text-retrieval and general machine learning ('hub'/'anti-hub' objects) where Radovanović et al. [1] were for the first time able to provide more general insight. They linked the problem to the property of distance/similarity score 'concentration' which occurs as a natural consequence of high dimensionality.

2.1. Concentration

Concentration is the property of a similarity or distance measure that corresponds to the tendency of objects in high dimensional data spaces to be almost equally distant from each other [7]. This results in pairwise distances to become almost identical to each other as the dimensionality increases, thus making it difficult to distinguish between the farthest and closest object. The effect of distance concentration has been proven for Euclidean spaces and other ℓ^p norms [7, 8]. Karydis et al. [9] empirically verified that the effect of distance

concentration can also be observed in music information retrieval applications computing spectral similarity. Concentration is usually measured as a ratio between some measure of spread and magnitude. For example, take the ratio between the standard deviation of all distances to an arbitrary reference point in a random vector space and the mean of these distances. If this ratio converges to zero as the data dimensionality goes to infinity, the distances are said to concentrate. In the case of the Euclidean distance and growing dimensionality, the standard deviation of distances converges to a constant while the mean keeps growing. Thus the ratio converges to zero and the distances are said to concentrate.

2.2. Hubs

Due to the phenomenon of concentration two things are now expected to happen in any i.i.d. random high-dimensional data distribution in the finite case [1]: (a) Points closer to the data center are *expected* to appear, which in turn have on average a higher similarity score to all other points. Due to this property these points, called hubs, keep appearing in the nearest neighbor lists of a lot of other objects. (b) Likewise, points which are farther away from the data center will emerge, which have a very low similarity score compared to all other data points. In the extreme case, these points will never occur in the nearest neighbor lists of other objects, and thus are called anti-hubs.

The emergence of hubs in the scores of a speaker verification system would also explain the emergence of the animals in the Doddington zoo, which is a categorization of objects as animals according to their score distribution and induced error. Hubs appear as wolves and lambs (impostors which have a high similarity score to all speaker models and vice versa) and anti-hubs appear as goats (impostors which have a small similarity score to all speaker models and thus are missed in the identification).

2.3. Score/Cohort Normalization

To alleviate the Doddington zoo effects, speaker verification systems employ normalization methods [4]. The multitude of normalization techniques can be divided into two major groups: 'cohort' and 'score' normalization methods. 'Cohort' normalization methods select a set of impostor speakers to normalize the similarity scores. The speakers are usually selected according to an acoustically motivated heuristic. e.g. as 'background speakers' by Reynolds [10]. 'Score' normalization techniques operate directly on the similarity scores and are applied to equalize inter-speaker variability and score distributions. The most broadly used score normalizing methods are the Z-Norm [11] or the T-Norm [12].

A closely related observation, that normalization techniques can be used to reduce the impact of hubs, has recently been made in the domain of general machine learning. Schnitzer et al. [13] introduced Mutual Proximity, a probabilistic distance normalization method which lessens the problem of hubs in general distance spaces, while at the same time quality and clustering criteria of the distance space improve. In the next section we will use this general purpose normalization method side by side with the methods used in verification systems to show that they have similar effects, although having been designed for different applications.

3. INVESTIGATION OF A SPEAKER VERIFICATION SYSTEM

In what follows we will investigate in three experiments (i) if the negative log-likelihood scores of a speaker verification system in fact concentrate in high dimensional feature spaces, (ii) if this leads to the emergence of hubs (and the animals in the Doddington zoo), and (iii) what the impact of normalization techniques in a speaker verification system is.

3.1. Speaker Verification

For our evaluations we build a system exactly replicating the setup and classic system described by Reynolds [10]. A single speaker model is represented as a GMM trained on the Mel frequency cepstrum coefficient (MFCC) representation of a number of training utterances. The GMM consists of 32 diagonal Gaussian components, estimated from a 2—40 dimensional MFCC representation. Before computing the MFCCs the speech signal is enhanced by spectral subtraction. A voice activity detector is then used to select the training frames. We use the NTIMIT [14] and TIMIT [15] databases, both containing 10 speech samples from 630 speakers each. These are smaller compared to the recent NIST SRE collections, but it has been observed in image and text-retrieval that hubness effects occur with increasing feature dimensionality regardless of collection size. Eight training samples are used to compute a speaker model (32 mixtures, diagonal covariance), the remaining two are used as impostor test samples. To identify speakers, the system computes the log-likelihood of a test sample given a speaker model. If the log-likelihood is higher than the identification threshold, the speaker was successfully identified.

Although the presented method alone is not state of the art, it is still an important component in recent speaker verification systems. Systems participating in the NIST 2010 speaker recognition evaluation such as the MIT LL system [16] or the SRI system [17] both still use GMMs as an integral part. To allow replication and reuse of the experiments on any dataset, the source code of the scripts is available online.[1]

3.2. Concentration of the Negative Log-Likelihood Scores

In our first experiment we measure if the similarity scores in the speaker verification system indeed concentrate when

[1] http://ofai.at/research/impml/projects/hubology.html

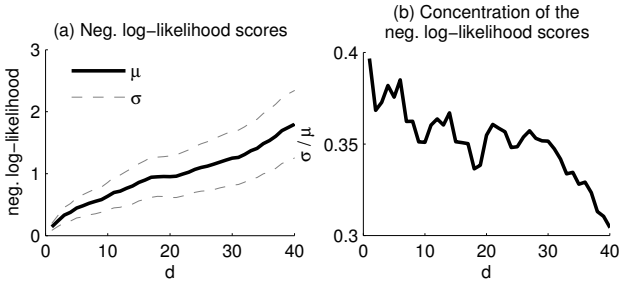

Fig. 1. Concentration effect observed in the negative log-likelihood scores of a speaker verification system with increasing feature dimensionality.

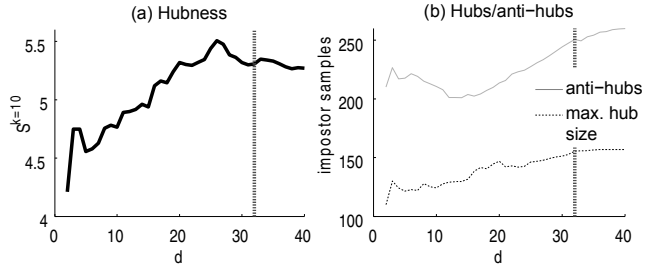

Fig. 2. Increase of hubness, number of hubs/anti-hubs with increasing feature dimensionality.

increasing the feature dimensions. We use the verification system described in the previous section together with the NTIMIT test database containing 168 speakers (1680 samples).

To measure the concentration of the similarity score space we compute the neighbor-range for each speaker model as it was done in [9]. The neighbor-range is computed for each speaker-model as the difference between the negative log-likelihood of the closest and k'th most similar impostor test sample. We compute the mean (μ) and standard deviation (σ) across all speaker-models at neighborhood ranges of $k = 10$.

To see the effect develop with increasing dimensionality, we start with 2 MFCCs and increase the number of dimensions gradually up to using 40 MFCCs as input to the GMMs. In each step we measure the concentration of distances as a ratio of σ and μ. We repeat the experiment five times and average the results. In each of the five iterations two different speaker samples are used as impostors, so that each sample is used as an impostor sample exactly once.

Figure 1a shows the increase of the mean negative log-likelihood and the measured standard deviation. While the measured mean score increases from 0.1 to 2, the standard deviation only slowly increases with higher feature dimensions. As the mean grows faster than the standard deviation, a score concentration effect (σ/μ) can be observed in Figure 1b.

3.3. Hubs in Speaker Verification Systems

The observation of a score concentration effect in high dimensions leads to the second experiment, an analysis of hubs in the examined speaker verification system. To measure the strength of the hubs phenomenon in any data base, Radovanović et al. [1] define a 'hubness' measure. To compute the measure, first the k-occurrence of each object x in a database is computed (N_x^k). N_x^k counts the number of times x occurs in the k nearest neighbor lists of all other objects in the collection. Hubness S^k is then defined as the skewness of the distribution of k-occurrences N^k,

$$S^k = \mathrm{E}\left[(N^k - \mu_{N^k})^3\right] \frac{1}{\sigma_{N^k}^3}.$$

Values close to zero indicate low hubness, high positive skewness indicates high hubness. To use these measures in a speaker verification system we first compute the similarity between all impostor test samples and speaker models. With this information we are then able to determine the k most similar impostor samples (nearest neighbors) of all speaker models and compute S^k.

We replicate the setup of the previous experiment and increase the number of MFCC dimensions of our speaker models step by step to measure hubness and the emergence of hub and anti-hub objects in the NTIMIT test database (168 speakers, 1680 samples). We compute the hubness $S^{k=10}$, the size of the biggest hub ($\max N^{k=10}$) and the number of objects which never occur as one of the most similar samples to any speaker model ($N^{k=10} = 0$). The results are again averaged over five runs.

Figure 2a plots the measured hubness in our experimental setup which is clearly increasing with growing feature dimensionality. Comparing this to Figure 1b, we can see that there is a simultaneous higher concentration of the similarity scores. At the same time we can see the size of the largest hub object, and number of anti-hubs increase in Figure 2b. Note the vertical dotted line at $d = 32$ in the two plots. At this point we can see that there is an impostor sample which occurs as close match ($k = 10$) in all of the 168 possible speaker models. At the same time over 250 samples never occur in the nearest neighbor lists. Both effects, one leading to wolves, the other to goats are undesirable effects and seem to strengthen with increasing feature dimensionality. At $d = 32$ we also reach the maximum upper limit of the maximal hub size since there already exists a hub appearing in all neighborhoods (Figure 2b). This effect also limits the measured skewness values (Figure 2a). In larger collections these measures have higher limits, see for example Section 3.5 (Table 1).

3.4. Normalization in Speaker Verification Systems

In our third experiment we measure the impact of normalization methods on hubs. We repeat the previous two experiments but apply normalization methods to the scores before

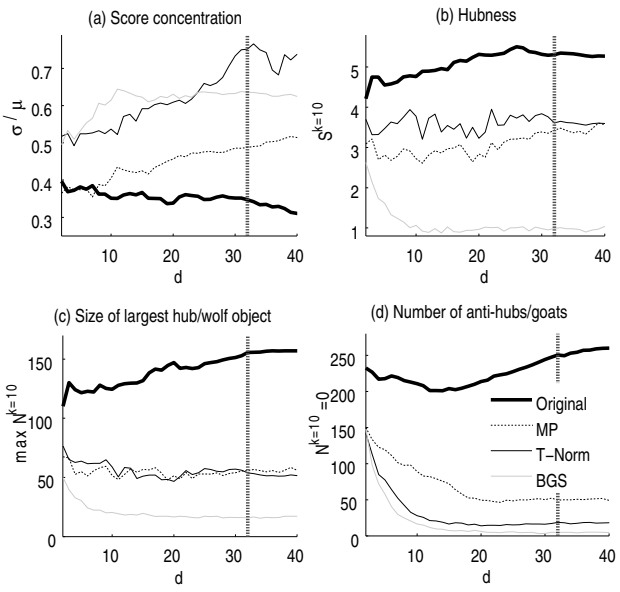

Fig. 3. The impact of normalization methods on hub related measures (bcd) and score concentration (a) compared to the original scores.

measuring the concentration and hubness (again using the NTIMIT test database, 168 speakers/1680 samples). We use background speaker (BGS) normalization [10] as representative for cohort normalization, T-Norm [12] as a representative for score normalization, and Mutual Proximity (MP) [13] as a method from general machine learning, unrelated to speaker verification systems.

Figure 3 collects the results of this experiment. In Figures 3bcd we can see that all measures related to hubs are greatly reduced when using a normalization method. For BGS the reduction in hubness is the greatest and yields the most balanced neighborhood (with a hubness of around 1, see Figure 3b). In terms of hubness, MP and T-Norm seem to perform similarly, reducing the hubness from 5.5 to 3.5. With clearly reduced hubness, the measured size of the largest hub object (Figure 3c) and the number of anti-hubs (Figure 3d) also decrease. The extreme values measured in the previous experiments are lowered significantly: the largest hub object occurs only 10 (BGS) or 60 (MP, T-Norm) times as nearest neighbor, while the number of anti-hubs are reduced to 60 (MP) or almost zero (T-Norm, BGS). We observe that the normalization method MP designed to alleviate the hub problem in general problem domains has similar effects like the domain specific normalization methods.

An interesting observation can be made when looking at Figure 3a where we compare the concentration of distance in the original negative log-likelihoods to the new normalized scores (MP, BGS, T-Norm). Apparently all normalization methods seem to de-concentrate the original scores, which in turn explains the reduced hubness, if we reverse the argumentation of Radovanović et al. [1].

Database	*Benchmark*	*Variants*			
		Original	BGS	MP	T-Norm
TIMIT	*EER*	**9.93**	**1.82**	**1.99**	**0.82**
	$S^{k=10}$	8.41	2.28	6.03	5.78
	$\max N^{k=10}$	347	35	128	106
	$N^{k=10}=0$	589	8	65	13
NTIMIT	*EER*	**20.21**	**8.59**	**7.38**	**4.38**
	$S^{k=10}$	11.46	2.67	8.25	9.58
	$\max N^{k=10}$	532	45	193	248
	$N^{k=10}=0$	914	70	213	187

Table 1. The impact of normalization on hubness and speaker verification system performance measured with the equal error rate (EER), hubness $S^{k=10}$, the size of the biggest hub ($\max N^{k=10}$) and the number of anti-hubs ($N^{k=10}=0$).

3.5. Impact on the Speaker Verification Performance

Until now we have only investigated the impact of the normalization methods on hubs by increasing the dimensionality of the features. In an additional experiment we survey the performance of the speaker verification system (using 20 dimensional MFCCs) in terms of the hub benchmark numbers and include the equal error rates (EER) to compare the quality of the systems. We use the complete TIMIT and NTIMIT databases (630 speakers and 1260 test samples each).

Table 1 shows the result of the evaluation. Like in our previous evaluations we observe extremely high hubness and a high number of hubs and anti-hubs in the original scores, with high EER and low system performance. Looking at the results for NTIMIT, we see a hubness ($S^{k=10}$) of 11.46 and the most severe hub sample ($\max N^{k=10}$) appearing in the neighborhood of 532 speaker models out of 630 possible. At the same time 914 speaker samples out of 1260 possible are anti-hubs ($N^{k=10}=0$) and are never close to any of the speaker models.

When using a normalization method, the measured error rates improve substantially while all hub benchmarks improve too. For NTIMIT the EER decreases significantly from 20.21 in the original space to 8.59, 7.38 and 4.38 depending on the normalization method used (BGS, MP, T-Norm). Anti-hubs ($N^{k=10}=0$) decrease from 914 down to 70, 213 and 187 and the size of the largest hub object ($\max N^{k=10}$) decreases from from 532 to 45, 193 and 248. This improvement can also be seen in the hubness values ($S^{k=10}$), although the values for normalization via MP and T-Norm remain somewhat high (8.25 and 9.58). Inspection of the distribution of k-occurrences N^k reveals that in both cases a single object with high k-occurrence is responsible for this remaining skewness, suggesting that the S^k measure is not an ideal measure of hubness alone. The number of hubs and anti-hubs should always be taken into consideration as well, as it has been done

throughout this work. Another important observation is the fact that T-Norm normalization, although achieving the best EER rate, is not as good at decreasing hubness as e.g. BGS. This tells us that hubness is an important part of explaining errors in speaker verification, but most likely not the only one.

Similar effects can be observed for the TIMIT set but are not described here in detail for lack of space. While the quality in terms of the EER varies by normalization method, all methods improve EER rates while reducing the number of hubs and anti-hubs.

4. DISCUSSION AND SUMMARY

We have investigated the relation of the different animals in the Doddington zoo in a speaker verification system to the concentration of distances and the hubness phenomenon. We have demonstrated that the higher the feature dimension, the more pronounced and problematic the effects become. As an implication for speaker verification systems, impostors which are able to impersonate a lot of other persons ('wolves') and audio samples which can never be identified correctly ('goats') are *expected* to emerge naturally as a consequence of high dimensionality. The insights gained from our experiments make us confident that hubness is not the only but an important part of explaining the emergence of the Doddington zoo.

This first link between two previously seemingly unrelated topics, hubs and the Doddington zoo, should trigger more efforts to understand the role of hubness in speaker verification. Directions for such future work are corroboration of our results on larger speaker data bases and using a more state-of-the-art speaker verification system. Increasing the dimensionality of the system by increasing the number of mixtures in the GMMs instead of increasing the number of features is another interesting aspect. And last but not least, a hubness aware normalization procedure designed specifically for speaker verification might be able to achieve even better system performance.

ACKNOWLEDGMENTS

This research is supported by the Austrian Science Fund (FWF): P24095. The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry for Transport, Innovation, and Technology.

5. REFERENCES

[1] M. Radovanović, A. Nanopoulos, and M. Ivanović, "Hubs in space: Popular nearest neighbors in high-dimensional data," *Journal of Machine Learning Research*, vol. 11, pp. 2487–2531, December 2010.

[2] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation," Tech. Rep., DTIC Document, 1998.

[3] N. Yager and T. Dunstone, "The biometric menagerie," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 220–230, 2010.

[4] N. Poh and M. Tistarelli, "Customizing biometric authentication systems via discriminative score calibration," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2681–2686.

[5] H. Jegou, C. Schmid, H. Harzallah, and J. Verbeek, "Accurate image search using the contextual dissimilarity measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 2–11, 2010.

[6] J.J. Aucouturier and F. Pachet, "A scale-free distribution of false positives for a large class of audio similarity measures," *Pattern Recognition*, vol. 41, no. 1, pp. 272–284, 2008.

[7] D. François, V. Wertz, and M. Verleysen, "The concentration of fractional distances," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, pp. 873–886, 2007.

[8] C. Aggarwal, A. Hinneburg, and D. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *Database Theory—ICDT 2001*, Lecture Notes in Computer Science, pp. 420–434. Springer Berlin/Heidelberg, 2001.

[9] I. Karydis, M. Radovanovic, A. Nanopoulos, and M. Ivanovic, "Looking through the glass ceiling: A conceptual framework for the problems of spectral similarity," in *Proceedings of the International Society for Music Information Retrieval (ISMIR10) Conference*, 2010.

[10] D.A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech communication*, vol. 17, no. 1, pp. 91–108, 1995.

[11] K.P. Li and J.E. Porter, "Normalizations and selection of speech segments for speaker recognition scoring," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 1988, pp. 595–598.

[12] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1, pp. 42–54, 2000.

[13] D. Schnitzer, A. Flexer, M. Schedl, and Widmer G., "Local and global scaling reduce hubs in space," *Journal of Machine Learning Research*, vol. 13, pp. 2813–2844, October 2012.

[14] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "Ntimit: A phonetically balanced, continuous speech, telephone bandwidth speech database," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP-90)*. IEEE, 1990, pp. 109–112.

[15] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[16] D. Sturim, W. Campbell, N. Dehak, Z. Karam, A. McCree, D. Reynolds, F. Richardson, P. Torres-Carrasquillo, and S. Shum, "The MIT LL 2010 speaker recognition evaluation system: Scalable language-independent speaker recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5272–5275.

[17] N. Scheffer, L. Ferrer, M. Graciarena, S. Kajarekar, E. Shriberg, and A. Stolcke, "The SRI NIST 2010 speaker recognition evaluation system," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5292–5295.