Interspeech 2022
18-22 September 2022, Incheon, Korea

Reliability criterion based on learning-phase entropy for speaker recognition with neural network

Pierre-Michel Bousquet, Mickael Rouvier, Jean-François Bonastre

LIA - Avignon University
first.lastname@univ-avignon.fr

Abstract

The reliability of Automatic Speaker Recognition (SR) is of the utmost importance for real-world applications. Even if SR systems obtain spectacular performance during evaluation campaigns, several studies have shown the limits and shortcomings of these systems. Reliability first means knowing where and when a system is performing as expected and a research effort is devoted to building confidence measures, by scanning input signals, representations or output scores. Here, a new reliability criterion is presented, dedicated to the latest SR systems based on deep neural network (DNN). The proposed approach uses the set of anchor speakers that controls the learning phase and takes advantage of the structure of the network itself, in order to derive a criterion making it possible to better assess the reliability of the decision based on the extracted speaker embeddings. The relevance and effectiveness of the proposed confidence measure are tested and demonstrated on widely used datasets.
Index Terms: Speaker recognition, speaker embeddings, x-vectors, reliability

1. Introduction

In recent years, Automatic Speaker Recognition (SR) has drawn on probabilistic or discriminative approaches to provide low-size total variability factors (now called speaker embeddings), equated with representations of statements directed towards the targeted task. The i-vector paradigm [1] relies on a solid theoretical basis, the probabilistic factor analysis framework, to provide such a representation. By opposition, the recent DNN-based approaches [2, 3, 4] use a task-driven strategy to learn the representation model. The loss function to be minimized during the training phase is the cross-entropy between the speakers present in the training set (we call them anchor speakers in the rest of this article). Then, the speaker-oriented representation of the test utterance is assimilated to the penultimate layer of the network (removing the output layer, which stores the entropy values of the loss function). This speaker embedding, usually referred to as x-vector, is then directly used to compute the scores, using probabilistic (PLDA [5]) or geometric (cosine [6]) metrics. Thus, the x-vector training is optimized following a close-set speaker classification task when a scoring approach such as PLDA is optimized to separate two classes, introducing a gap between the objectives of these two phases.

The reliability of the SR was addressed in several research projects, with interest often motivated by the quality requirements of forensic applications [7, 8, 9]. Objective performance benchmarking has also aroused much research [10, 11]. Methods have been proposed to provide confidence measures, combining different sources of information [12, 13, 14, 15]. Much work probe the quality of the signals involved in the speaker verification or the likelihood of the data given the model [16]. Dependency of the speaker specific information to the phonological content has been revealed as well as the influence of the speech rhythm variability [17, 18, 19].

About modeling and scoring, the weakness of the decision can be due to the lack of discriminative information in the representations [20] or to mismatch between model metaparameters and the voice records [21]. Some approaches rely on Bayesian networks to obtain a probabilistic measure of the reliability of the trial [22, 23, 24, 25]. Many other solutions have been tested, among which vector Taylor series (VTS) [16] and information theory [26, 27]. To our knowledge there are no studies devoted to the reliability of speaker embeddings extracted by neural network, in terms of goodness-of-fit.

Here, a new type of confidence measure is proposed for this purpose. Its originality is to directly take into account the gap between the objectives of the training and scoring phase. We hypothesize that this bias could imply an implicit -and therefore undetectable- bias within the x-vectors. Starting from this assumption, we design a specific SR reliability criterion based on the compliance of the DNN model (observed during the x-vector extraction phase). The proposed solution is based on local goodness-of-fit measures, computed during the learning phase.

2. Compliance with model and feature reliability

2.1. Learning-phase entropy

From the designer's and user's point of view, the goal of a DNN for SR is to provide a fixed-size vector-representation of an utterance, fitted to the speaker discrimination task. From the "machine" point of view, the goal is to estimate a function that relates the parameters of the acoustic signal to a speaker identity vector. The motivation of our investigation is illustrated in Figure 1-a, where the output vector p (the last layer of the trunk architecture) for an example of the training speaker s_k is computed and the value $-\log(p_k)$ is added to the DNN loss function. For example, with the angular softmax loss function [29, 30], p is equal to:

$$p = \left[\frac{e^{s(\cos(\mathbf{W}_i, x) - m_i)}}{\sum_j e^{s\left(\cos\left(\mathbf{W}_j, x\right) - m_j\right)}} \right]_{i=1,\ldots,n} \quad (1)$$

where n is the number of training speakers, x is the input to the layer (the x-vector), s is an adjustable scale factor, $\mathbf{W}$ is the weight matrix and m is the penalty margin, equal to 0 if $i \neq k$.

The fact that the second to last layer is selected then handled to provide PLDA or cosine based decisions is ignored during the learning phase: the system is only trained to well separate the anchor speakers of the training dataset. What about a new utterance, from a speaker who is unknown to the system? For us, the hope its extracted speaker embedding x will allow to

10.21437/Interspeech.2022-8

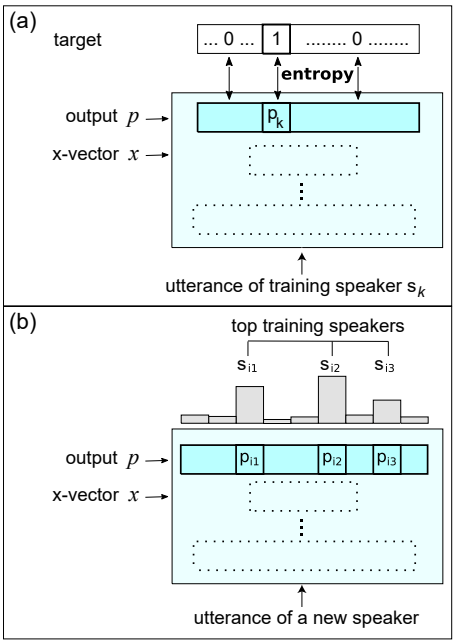

Figure 1: *Illustration of the DNN configuration stages and learning-phase components described in Section 2.1: (a) what happens during the DNN learning phase? (b) ... and when extracting features of new speakers?*

discriminate this speaker. For the machine, the system provides an output vector p, comprised of the values involved in the loss function (dropping the penalty margin). As shown in Figure 1-b, the model is telling us that this utterance can be classified as $p_{i_1}\%$ of speaker s_{i_1}, $p_{i_2}\%$ of speaker s_{i_2}, etc... Assuming that some values of p are dominant, a subset of *top*-training speakers can be determined for this utterance, that is of highest values p_k (in the figure: s_{i_1}, s_{i_2}, s_{i_3}). The method for determining the top-training speakers is detailed in Section 2.2.4.

The output vectors p of the examples of the training speakers dataset can be computed. During the learning phase these vectors are optimized to be as close as possible to the desired values, the target vector equal to 1 for the index k of the training speaker k, otherwise 0 (Figure 1-a). The shift between the example outputs and the desired target can be seen as the entropy of the DNN learning phase, in terms of compliance with model of a given training speaker. Moreover, this shift could contribute to reduce the gap (maximal on the target vectors) between one training speaker and another, which could lead to confusion between them. Although this entropy does not necessarily penalize the resulting x-vector of the new utterance, this is a *local loss* that has to be considered in a goodness-of-fit study. Therefore, the goal of our investigation is, first, to determine some locality sensitive goodness-of-fit measurements, based on information theory. Then to transmit these measures of nuisance, which are resulting from the lack of compliance with model during the DNN learning phase, to the data that are extracted. Lastly, a reliability criterion of a trial can be proposed, which combines these entropy levels.

2.2. Reliability criteria

2.2.1. Identification criterion

Given an example of the k^{th} training speaker s_k and its softmax-transformed output vector $p = [p_i]_{i=1}^n$, as defined for example in Eq. 1, the entropy of the target vector $\delta^{(k)}$, equal to 1 if k otherwise 0, a posteriori of p is equal to the Kullback-Leibler divergence $D_{KL}\left(\delta^{(k)}||p\right) = -\log\left(p_k\right)$. This is the part of the loss function induced by this example of s_k. Therefore, the level of compliance with model of this utterance is defined as $\log\left(p_k\right)$, and the first reliability criterion of a new utterance utt (of a speaker unknown to the system) is obtained by averaging on all the examples of all the top-training speakers for this utterance:

$$r_1\left(utt\right) = \frac{1}{n_{top}} \sum_{k \in top} \frac{1}{n_k} \sum_{p \in s_k} \log\left(p_k\right) \tag{2}$$

where top is the index subset of top-training speakers for this utterance, n_{top} their number and n_k the size of the k^{th} speaker's sample. It can be referred to as an *identification* criterion, as it measures the ability of the model to properly fit the training speakers mainly involved in the representation of the new utterance.

2.2.2. Discrimination criterion

The reliability of a new utterance is also depending on the ability of the model to properly discriminate between these top-training speakers. On the one hand, the previous measure does not take into account critical cases of high "foreign" values, that is, high p_i values, $i \neq k$ for the k^{th} training speaker. If the entropy of these foreign values is not maximal and, above all, when some of them are significant, this could lead to a risk of confusion between s_k and other anchor speakers.

To favor maximal entropy on the foreign values, the following criterion is defined, for each utterance χ of the training speaker s_k:

$$\begin{aligned} c\left(\chi\right) &= -D_{KL}\left(\left[p_i\right]_{i \neq k} ||\mathcal{U}\right) \\ &= -\left(\sum_{i \neq k} \frac{p_i}{\sum\limits_{j \neq k} p_j} \log\left(\frac{p_i}{\sum\limits_{j \neq k} p_j}\right) + \log\left(n-1\right)\right) \end{aligned} \tag{3}$$

where $D_{KL}\left(\left[p_i\right]_{i \neq k} ||\mathcal{U}\right)$ is the Kullback-Leibler divergence of $\left[p_i\right]_{i \neq k}$ a posteriori of the uniform distribution $\mathcal{U}$. The criterion is maximal when all the foreign values $i \neq k$ are equal. The second reliability criterion of a new utterance stems from this one, by averaging as done in Eq. 2:

$$r_2\left(utt\right) = \frac{1}{n_{top}} \sum_{k \in top} \frac{1}{n_k} \sum_{\chi \in s_k} c\left(\chi\right) \tag{4}$$

The result can be referred to as an *absolute* discrimination criterion.

2.2.3. Other discrimination criterion

On the other hand, to assess the ability to discriminate the k^{th} and l^{th} training speakers s_k and s_l, one can use the average Jeffreys divergence between their examples :

$$\mathbf{J}\left(k,l\right) = \frac{1}{n_k n_l} \sum_{p \in s_k} \sum_{q \in s_l} \left(D_{KL}\left(p||q\right) + D_{KL}\left(q||p\right)\right)$$

It stems from this compliance level the following discrimination criterion of a new utterance:

$$r_3\left(utt\right) = \frac{1}{n_{top}\left(n_{top}-1\right)} \sum_{k \in top} \sum_{\substack{l \in top \\ l \neq k}} \mathbf{J}\left(k,l\right) \qquad (5)$$

It can be referred to as a *relative* discrimination criterion.

2.2.4. Top-training speakers and similarity criterion

To determine the top-training speakers subset of an utterance, the index subset I of the highest p values is retained, such that $\sum_{k \in I} p_k$ exceeds a threshold α, that is $\alpha\%$ of the sum of p-values. In our experiments, α is set to 75%.

Moreover, some experiments have shown that the number of top-speakers of the test utterances is usually very low (less than 100 among 6000 training speakers). The network, trained as a classifier for the training set, proceeds with new speakers by similarity to it. Therefore, too many top-speakers could reveal some difficulties in the system to fit the data. The following *similarity* criterion between the model and a new utterance can be proposed:

$$r_4\left(utt\right) = -n_{top}\left(utt\right) \qquad (6)$$

that is, the number of top-training speakers for this utterance (with minus sign to have increasing reliability).

2.2.5. Final reliability criterion of a trial

To provide a unique reliability criterion for an SR trial, the four previous criteria are first normalized : given an utterance u, each value $r_i(u)$ is replaced by its quantile $R_i(u)$, i.e. the value such that $R_i(u)\%$ of the r_i are lower than $r_i(u)$. By this way, all the criteria lie in $[0, 1]$. The quantiles are estimated on a development set, then the values are computed for enrollment and test utterances and the proposed final reliability criterion of a trial $t = (enroll, test)$ is:

$$R\left(t\right) = \frac{1}{4}\sum_{i=1}^{4} \min\left(R_i\left(enroll\right), R_i\left(test\right)\right) \qquad (7)$$

The minimal values are retained to penalize more severely the weaknesses of the model.

3. Experiments

3.1. Experimental setup

The x-vector extractor used in this paper is a variant based on ResNet-34. The extractor was trained on the development part of the Voxceleb 2 dataset [31], cut into 4-second chunks and augmented with noise, as described in [3] and available as a part of the Kaldi-recipe. It contains about 1M segments (+ 4M augmented) of 5994 speakers. As input, we used 60-dimensional filter-banks. The speaker embeddings are 256-dimensional and the loss is the angular additive margin with scale equal to 30 and margin equal to 0.2. The sizes of the feature maps are 256, 256, 512 and 512 for the 4 ResNet blocks. We use stochastic gradient descent with momentum equal to 0.9, a weight decay equal to 2.10^{-4} and initial learning rate equal to 0.2. The implementation is based on PyTorch. For scoring, the x-vectors are centered by subtracting the overall mean of the training dataset, then the cosine metric is applied. For computing the compliance levels of the training speakers, only the original segments of VoxCeleb 2 are used (no augmented data).

The relevance of the reliability criterion is tested on six datasets: -VoxCeleb 1 [32], which is completely disjoint from the training dataset, -Speakers in the Wild (SITW) [33] with 4170 enrollment and 2275 test utterances, -the English part of Librispeech [34], a dataset derived from read audiobooks from LibriVox, -the dataset used for task 2 of the Short duration speaker verification (SDSV) challenge 2020, a text-independent speaker recognition evaluation based on the recently released DeepMine dataset [35, 36], comprised of Persian-native and some English-non native utterances, -Snips [37] an English dataset initially used to compare different voice assistants, -Cn-celeb [38], a speaker recognition dataset collected from Chinese celebrities.

For SITW, the trial set is the one of the core-core task. For all others, a large trial dataset is built by crossing the available data in such a way that the target probability is around 1%. Table 1 shows the sample sizes of speakers, utterances (enrollment+test) and trials, and the equal error rate (EER) obtained with our scoring.

Table 1: *Evaluation datasets designed for the experiments*

	#spk	#utt	#trials	EER (%)
VoxCeleb1	486	50 819	58 M	1.15
SITW	-	6 445	721 K	1.23
librispeech	560	44 800	101 M	0.94
SDSV	504	35 280	62 M	2.91
Snips	515	23 690	28 M	5.22
Cn-celeb	601	40 480	82 M	21.60

3.2. Analysis

For each evaluation, the trial dataset is split into 50 equally sized subsets, depending on intervals of the final reliability criterion R of Eq. 7. Then the equal error rate (EER) of each trial subset is computed. Figure 2 shows the curves of the EERs (y-values) by increasing reliability (x-values). The dashed line is the average EER computed on the overall set. All the curves are significantly decreasing. The uncertainty in the SITW curve can be explained by the small size of the core-core trial set (only 3658 positive trials). From the least to the most reliable trials, the EER is divided by about 2 for Snips and Cn-Celeb and by 3 to 8 for other evaluations. These results show the dependency of the SR accuracy on the local goodness-of-fit of the model.

Moreover, it can be noticed that this uncertainty is related to the accuracy of the system : the higher the latter, the higher the former. The fact that high accuracy can lead to greater relative uncertainty could be imputed to the sensitivity of the system (to the quality of the signal or to any data mismatch). It is proven here that this is also due to the learning phase entropy.

3.3. More in-depth analysis

The previous EERs are computed on evaluation subsets, of which the hypothesis of uniformity in terms of speaker could be rebutted. To address this concern, for each evaluation, the unique EER decision threshold computed on the overall trial dataset is applied to extract the trials detected as positive (same speaker), false or true. Then, for 50 equally sized intervals q of increasing reliability R, the number of false and true positive trials $n_{F,q}$ and $n_{T,q}$ into each corresponding subset is computed to provide the "False Positive rate" $n_{F,q}/(n_{F,q}+n_{T,q})$. Figure 3 shows the curve of this False Positive rate depending on R

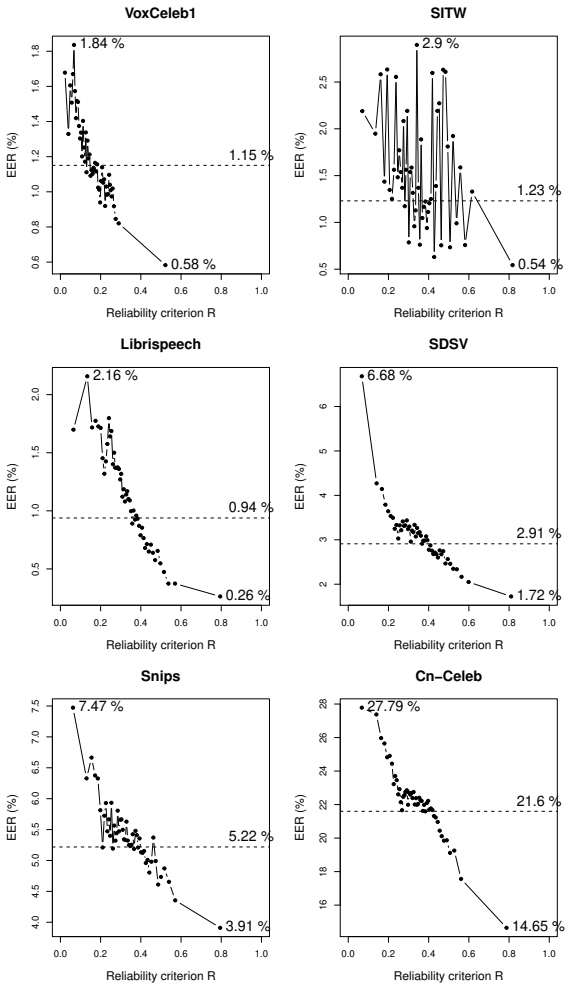

Figure 2: *Equal error rates of evaluation subsets depending on the final reliability criterion R of Eq. 7 (x-values into $[0, 1]$). Each point summarizes a fiftieth of the trial set. The dashed line is the EER computed on the overall set.*

(only two evaluations are displayed, due to lack of space). The values should all be equal to the overall prior (dashed line) if the criterion is ineffective. The dotted curves are the minimal and maximal value obtained by 10%-leave-out cross validation, to provide a confident interval for the series. The decreasing trends of the curves confirm the effectiveness and relevance of the proposed reliability criterion.

One could be pointed out that similar results would be achieved with the score instead of the proposed criterion: the higher the score (i.e. the further the score from the decision threshold), the most reliable the decision. This is only true if the model is supposed to be perfect in terms of goodness-of-fit. Table 2 reports correlation coefficients between score and reliability criterion for trials detected as positive. All correlations are close to 0, which confirms that the proposed uncertainty measure cannot be achieved by using the scoring metric: the reliability criterion measures the quality of the modeling for a trial while the score provides a decision a posteriori of a well-fitted model.

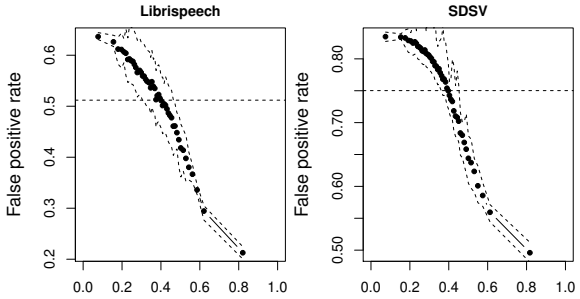

Figure 3: *False positive ratio, among trials detected as positive, depending on R intervals (x-values into $[0, 1]$). The values should all be equal to the prior (dashed line) if the criterion is ineffective.*

Table 2: *Correlation between score and reliability criterion R.*

VoxCeleb1	SITW	Librispeech	SDSV	Snips	Cn-celeb
0.15	0.16	0.16	0.08	0.05	0.05

4. Conclusion

In recent years, many confidence measures of Automatic Speaker Recognition (SR) system reliability have been proposed. They are usually based on the characteristics of the input speech extracts and the distribution of the output scores.

This study takes advantage of the specific configuration of the current leading approach in SR, the DNN-based systems. Given an utterance, the proposed approach probes the region of the model mainly involved in the estimation of the corresponding embedding. Using well-known measures of information theory, the method produces a reliability criterion for one trial in terms of local-wise goodness-of-fit. This criterion can be considered as a prerequisite indicator, essential before any in-depth reliability diagnosis focusing on the SR system inputs or outputs. The experiments proposed in this article show that the criterion can point out modeling weaknesses, which trigger a substantial inflation in detection errors. They also demonstrate that the uncertainty in the accuracy of SR systems is not only due to the quality of the speech signals or the shift between training and test data, but also to specificity of the modeling phase.

The criterion introduced here opens up perspectives for future work. It can initiate entropy-based training enhancement. The second intermediate criterion could also be included in the DNN cross-entropy loss function, to improve the discriminative training. Above all, the final reliability criterion allows for the evaluation of high security and forensic decisions (*acceptability* of the decision) and, therefore, needs to be improved and completed until fully assessing the impact of the modeling phase on the reliability of the decision.

5. Acknowledgements

This research was supported by the ANR agency (Agence Nationale de la Recherche), VoxCrim project (ANR-17-CE39-0016)

6. References

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE*

Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788–798, 2011.

[2] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 165–170.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *ICASSP*. IEEE, 2018, pp. 5329–5333.

[4] J. Rohdin, A. Silnova, M. Diez, O. Plchot, P. Matějka, and L. Burget, "End-to-end DNN based speaker recognition inspired by i-vector and PLDA," in *ICASSP*. IEEE, 2018, pp. 4874–4878.

[5] S. J. D. Prince, *Computer Vision: Models, Learning, and Inference*, 1st ed. New York, NY, USA: Cambridge University Press, 2012.

[6] D. Garcia-Romero, D. Snyder, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "x-Vector DNN Refinement with Full-Length Recordings for Speaker Recognition," in *Interspeech*, 2019, pp. 1493–1496.

[7] J. Campbell, W. Shen, W. Campbell, R. Schwartz, J. Bonastre, and D. Matrouf, "Forensic speaker recognition," *IEEE Signal Processing Magazine*, vol. 26,2, pp. 95–103, March 2009.

[8] G. S. Morrison, C. Zhang, and P. Rose, "An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system," *Forensic Science International*, vol. 208, no. 1, pp. 59–65, 2011.

[9] M. Ajili, "Reliability of voice comparison for forensic applications," Ph. D., Avignon University, 2017.

[10] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. A. Reynolds, "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation," in *ICSLP*. ISCA, 1998.

[11] N. Brummer, "Measuring, refining and calibrating speaker and language information extracted from speech," Ph.D. dissertation, Stellenbosch: University of Stellenbosch, 2010.

[12] W. M. Campbell, D. A. Reynolds, J. P. Campbell, and K. J. Brady, "Estimating and evaluating confidence for forensic speaker recognition," in *ICASSP*, vol. 1, 2005, pp. I/717–I/720 Vol. 1.

[13] R. Haraksim, D. Ramos, D. Meuwly, and C. E. Berger, "Measuring coherence of computer-assisted likelihood ratio methods," *Forensic Science International*, vol. 249, pp. 123–132, 2015.

[14] P. Rose, "Technical forensic speaker recognition: Evaluation, types and testing of evidence," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 159–191, 2006.

[15] G. S. Morrison, "Measuring the validity and reliability of forensic likelihood-ratio systems," *Science & Justice*, vol. 51, no. 3, pp. 91–98, 2011.

[16] J. Villalba, A. Ortega, A. Miguel, and E. Lleida, "Analysis of speech quality measures for the task of estimating the reliability of speaker verification decisions," *Speech Communication*, vol. 78, pp. 42–61, 2016.

[17] M. Antal and G. Toderean, "Speaker recognition and broad phonetic groups," in *ICSPPRA*, M. H. Hamza, Ed. IASTED/ACTA Press, 2006, pp. 155–159.

[18] M. Ajili, J.-F. Bonastre, W. B. Kheder, S. Rossato, and J. Kahn, "Phonological content impact on wrongful convictions in forensic voice comparison context," in *ICASSP*. IEEE, 2017, pp. 2147–2151.

[19] V. Dellwo, A. Leemann, and M.-J. Kolly, "Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors," *The Journal of the Acoustical Society of America*, vol. 137, no. 3, pp. 1513–1528, 2015.

[20] W. Rao and M.-W. Mak, "Boosting the performance of i-vector based speaker verification via utterance partitioning," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 5, pp. 1012–1022, 2013.

[21] J. Kahn, S. Rossato, and J.-F. Bonastre, "Intra-speaker variability effects on speaker verification performance," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2010.

[22] J. Richiardi, P. Prodanov, and A. Drygajlo, "A probabilistic measure of modality reliability in speaker verification," in *ICASSP*, vol. 1. IEEE, 2005.

[23] J. Richiardi, A. Drygajlo, and P. Prodanov, "Confidence and reliability measures in speaker verification," *Journal of the Franklin Institute*, vol. 343, no. 6, pp. 574–595, 2006.

[24] J. Villalba, E. Lleida, A. Ortega, and A. Miguel, "Reliability estimation of the speaker verification decisions using bayesian networks to combine information from multiple speech quality measures," in *Advances in Speech and Language Technologies for Iberian Languages*. Springer, 2012, pp. 1–10.

[25] ——, "A new bayesian network to assess the reliability of speaker verification," in *InterSpeech*, 2017.

[26] D. Ramos, J. Gonzalez-Rodriguez, G. Zadora, and C. Aitken, "Information-theoretical assessment of the performance of likelihood ratio computation methods," *Journal of forensic sciences*, vol. 58, no. 6, pp. 1503–1518, 2013.

[27] A. Nautsch, C. Rathgeb, R. Saeidi, and C. Busch, "Entropy analysis of i-vector feature spaces in duration-sensitive speaker recognition," in *ICASSP*. IEEE, 2015, pp. 4674–4678.

[28] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," *CoRR*, vol. abs/1704.08063, 2017.

[29] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[30] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," 2019.

[31] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *Interspeech 2018*, Sep 2018.

[32] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Interspeech 2017*, Aug 2017.

[33] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (SITW) speaker recognition database." in *Interspeech*, 2016, pp. 818–822.

[34] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[35] H. Zeinali, H. Sameti, and T. Stafylakis, "DeepMine speech processing database: Text-dependent and independent speaker verification and speech recognition in Persian and English." in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2018, pp. 386–392.

[36] H. Zeinali, L. Burget, and J. Cernocky, "A multi purpose and large scale speech corpus in Persian and English for speaker and speech recognition: the DeepMine database," in *Proc. ASRU 2019 The 2019 IEEE Automatic Speech Recognition and Understanding Workshop*, 2019.

[37] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril, M. Primet, and J. Dureau, "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces," *CoRR*, vol. abs/1805.10190, 2018.

[38] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "Cn-celeb: a challenging chinese speaker recognition dataset," 2019.