# Speech database and protocol validation using waveform entropy

*Itshak Lapidot*[1], *Héctor Delgado*[2], *Massimiliano Todisco*[2],
*Nicholas Evans*[2] *and Jean-François Bonastre*[3]

[1]Afeka Tel-Aviv College of Engineering, ACLP, Israel
[2]EURECOM, Biot France
[3]University of Avignon, LIA, France

itshakl@afeka.ac.il, hector.delgado@eurecom.fr, Massimiliano.Todisco@eurecom.fr,
evans@eurecom.fr, jean-francois.bonastre@univ-avignon.fr

## Abstract

The assessment of performance for any number of speech processing tasks calls for the use of a suitably large, representative dataset. Dataset design is crucial so as to ensure that any significant variation unrelated to the task in hand is adequately normalised or marginalised. Most datasets are partitioned into training, development and evaluation subsets. Depending on the task, the nature of these three subsets should normally be close to identical. With speech signals being subject to a multitude of different influences, e.g. speaker gender and age, language, dialect, utterance length, etc., the design and validation of speech datasets can become especially challenging. Even if many sources of variation unrelated to the task in hand can easily be marginalised, other sources of more subtle variation can easily be overlooked. Imbalances between training, development and evaluation partitions, can bring into question findings derived from their use. Stringent dataset validation procedures are required. This paper reports a particularly straightforward approach to dataset validation that is based upon waveform entropy.

**Index Terms**: Database assessment, waveform, entropy

## 1. Introduction

The assessment of any statistical pattern recognition algorithm calls for experimentation using suitably large and representative datasets. Speech tasks such as automatic speech recognition (ASR) [1,2] and automatic speaker verification (ASV) [3,4] are no exception. On the contrary, speech domain is one area where the evaluation paradigm takes a large place. The story started from DARPA program and the number of evaluation campaigns or challenges increases year after year [5–9]. If the interest of such an evaluation paradigm is well established, a badly defined protocol could have an important impact on the development of one scientific area or important consequences on the citizens, for example when forensic aspects are taken into consideration.

While the design of datasets to support ASR and ASV research may not at first appear complicated, it is crucial that careful attention be paid so as to ensure that any significant variation unrelated to the task is adequately normalised or marginalised. For example, the assessment of ASR and ASV systems may call for the use of a dataset with an acceptable speaker gender or age balance that is representative of the foreseen real-world application, e.g. telephone banking. Other tasks may call for the use of datasets that capture language and dialect variation, e.g. surveillance.

Whatever the dataset and whatever the application, there is generally a need to partition the database into a number of subsets in order to facilitate experimentation. So-called protocols with at least three *independent* partitions are common (where the notion of independence is application dependent). A training partition may be needed to support the learning of background information, e.g. the universal background model (UBM) used in ASV [3]. A development partition can be used for system optimisation whereas an evaluation partition is generally used to assess system generalisation to new data. The use of a fourth validation partition is sometimes desirable. Be there three or four partitions, the objective is to obtain a reliable *estimation* or *prediction* of recognition or classification performance – what could be expected were the system to be deployed in the wild.

The reliability of the performance prediction is critically dependent upon the quality of the dataset and of the protocols. Speech signals are inherently dynamic in nature and reflecting both biological and behavioural influences [10, 11], in addition to multiple sources of variation, e.g. utterance length, health, emotion, in addition to those mentioned above (and many more). The potential impact of such variation upon assessment must be addressed, ideally before dataset collection and, crucially, for protocol design. Otherwise, the inadequate marginalisation of nuisance variation and imbalances between training, development and evaluation partitions can jeopardise the integrity of results and findings. Stringent dataset and protocol validation procedures are thus mandatory.

While it is certainly likely that a robust approach to dataset and protocol validation should encompass a battery of different measures and procedures, this paper proposes one such approach based on measurements of waveform entropy. The approach is straightforward and efficient and can be applied on-the-fly as part of protocol design. The entropy validation tool provides a visualisation of the entropy distribution in the form of a *probability mass function* (PMF). Its application to training, development and evaluation partitions can be used for the rapid inspection of data partition imbalances.

The remainder of the paper is organized as follows. The waveform entropy validation tool is detailed in Section 2. Section 3 describes the databases to which the validation tool is applied. They include the TIMIT database, the RSR 2015

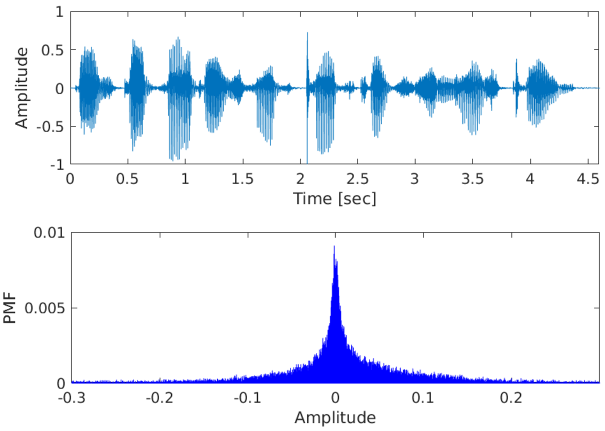

Figure 1: *An arbitrary bona fide speech utterance from the ASVspoof 2017 database and its corresponding signal amplitude distribution.*

database, the ASVspoof 2015 database and the ASVspoof 2017 v2.0 database. Results are described in Section 4. Conclusions are presented in Section 5. They discuss the significance of the findings and show the importance of adequate database and protocol validation.

## 2. Entropy validation

This section describes the proposed approach to database and protocol validation. It is based upon estimates and the resulting distribution of waveform entropy.

The entropy for a given speech signal is determined from the distribution of waveform amplitude given a particular coding or quantization. This is achieved by simply counting the occurrence of each quantisation level, thereby giving the speech signal amplitude distribution. Assuming a speech signal for which the amplitude of each sample is represented by 16 bits per sample, then there are $2^{16}$ possible quantisation levels. The histogram of waveform amplitude is determined by dividing the bin counts by the number of samples in the signal, thereby giving the signal amplitude *probability mass function* (PMF). The entropy of the *signal amplitude* PMF is then determined according to the standard approach given by:

$$H\left(s\right) = -\sum_{P_k \neq 0} P_k \log_2 P_k; \quad k \in \left\{1, \ldots, 2^{16}\right\} \tag{1}$$

where $H(s)$ denotes the waveform entropy for signal $s$, $P_k$ denotes the normalised histogram bin count and where $k$ is the histogram bin index. An example speech signal and corresponding amplitude distribution is illustrated in Fig. 1. This operation is performed separately for each speech signal $s$ in a given dataset. A *dataset entropy* PMF is then determined from the distribution of waveform entropies.

This procedure may be applied to all recordings in an entire dataset, or separately to distinct data partitions such as the training, development and evaluation subsets. When applied separately, the entropy validation tool, albeit a straightforward means of analysis, provides a rapid visualisation of data subset similarity or consistency in terms of a PMF. While not reported in this paper, the similarity between PMFs for two different data subsets could be expressed quantitatively, e.g., via the Kullback–Leibler divergence.

Depending on the application, differences between the entropy PMFs for two different datasets can serve to validate the integrity of the data partition (in the case that the entropy PMFs should be similar) and, alternatively, it can serve as a means of gauging task difficulty (in the case that differences are expected by design). Examples of both cases follow later in this paper.

## 3. Databases

Experiments reported in this paper relate to four standard databases, namely the TIMIT database [12], the RSR 2015 database [13], the ASVspoof 2015 database [6] and the ASVspoof 2017 v2.0 database [7, 14]. Statistics of each, including the number of speakers and number of speech segments in training, development and evaluation partitions are illustrated in Table 1. The four databases are described in the following.

### 3.1. Databases

**TIMIT.** This database was primarily designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems, although it has been used in many research works. It contains 16 kHz read-speech recordings captured in a noise-isolated recording booth using a high-quality headset microphone. As all the recordings were made in the same isolated conditions and using same recording device, the expectation is that the PMFs of the training the evaluation datasets should be close to identical.

**RSR 2015.** This database was developed to support research in text-dependent automatic speaker recognition. It contains read-speech recordings captured using 6 different handset devices. Three of these dominate the development set while the remaining three dominate the evaluation set. Channel differences are expected to create differences in the dataset entropy PMFs.

**ASVspoof 2015.** This database supports research in anti-spoofing. The task is to distinguish between bona fide (genuine speech) and spoofed speech produced with speech synthesis and voice conversion algorithms. Bona fide samples comprise high-quality, clean recordings captured in a semi-anechoic room with a solid floor. They correspond to the VCTK corpus[1]. Spoofed speech is generated according to ten (S1-S10) different speech synthesis and voice conversion algorithms. Since bona fide speech was collected in similar conditions, dataset entropy PMFs are expected to be consistent across all dataset partitions. Spoofed speech in training and development partitions share the same conditions S1-S5 (speech synthesis / voice conversion algorithms), while an addition set of speech synthesis / voice conversion algorithms, S6-S10, were used in the creation of the evaluation set. For this reason, dataset entropy PMFs for the evaluation set are expected to show soem variation compared to

[1] http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html

Table 1: *Stastistics of the TIMIT, RSR 2015, ASVspoof 2015 and ASVspoof 2017 databases.*

| | | TIMIT | RSR 2015 | ASVspoof 2015 | ASVspoof 2017 |
|---|---|---|---|---|---|
| #Speakers | Train | 326 / 136 | - | 10 / 15 | 10 |
| (male / female) | Development | - | 50 / 47 | 15 / 20 | 8 |
| | Evaluation | 112 / 56 | 57 / 49 | 20 / 26 | 24 |
| #Segments | Train | 4,620 | - | 3,750 / 12,625 | 1,507 / 1,507 |
| (bona fide / spoof) | Development | - | 63,640 | 3,497 / 49,875 | 760 / 950 |
| | Evaluation | 1,681 | 69,603 | 9,404 / 184,000 | 1,298 / 12,008 |

the training and development subsets.

**ASVspoof 2017 V2.** This database was created to support research in spoofing detection for replay spoofing attacks. Bona fide speech recordings, corresponding to the RedDots base corpora[2], include a substantial variation in recording devices and background noise. Replay spoofing recordings reflect heterogeneous recording and playback conditions, using devices of varying quality in acoustic environments with highly variable noise. Since the ASVspofo 2017 v2.0 database reflects substantial variability in terms of environmental conditions, recording devices and replaying methods, it likely that different subsets will exhibit significant variation in terms of dataset entropy PMFs. The dataset is unlikely to be sufficiently large such that the variation is fully balanced.

### 3.2. Pre-processing

Prior to entropy analysis, some datasets were treated with voice activity detection (VAD). For ASV tasks, discriminant information is contained only in intervals containing speech. For spoofing detection tasks, there is discriminant information in both speech and non-speech intervals. Accordingly, VAD was applied to TIMIT and RSR datasets, but not to the ASVspoof datasets.

VAD is performed according to approach described in [15] and [16]. Every $100ms$ a norm-2 signal was calculated according to $E_n = \left[\sum_{k=0}^{1599} s^2\left(1600n + k\right)\right]^{0.5}$. A threshold was defined as $Th = \alpha\left(\max\left\{E_n\right\} - \min\left\{E_n\right\}\right) + \min\left\{E_n\right\}$ and only frames above the threshold were used for entropy calculation. All experiments reported in this paper correspond to a value of $alpha = 0.03$.

## 4. Results

Presented here are results for a set of dataset and protocol validation experiments performed using the datasets described in Section 3.

### 4.1. TIMIT

Entropy validation results for the TIMIT database are illustrated in Fig. 2. Separate dataset entropy PMFs are shown for the standard training and development partitions. The two PMFs are largely overlapping suggesting that the two datasets are similar in nature, both in terms of speakers and acoustic content related to the read sentences that compose the TIMIT database. In view of the applications for which the TIMIT dataset was collected, it would appear that the dataset partition and protocol are well designed.

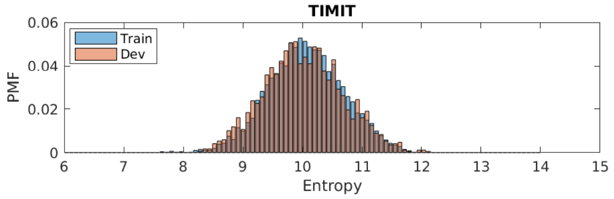

Figure 2: *TIMIT dataset entropy PMFs for training and development subsets.*

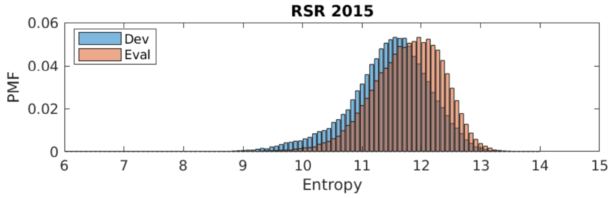

Figure 3: *As for Fig. 2 except for the RSR 2015 database.*

### 4.2. RSR 2015

Results for the RSR 2015 database are illustrated in Fig. 3. Separate dataset entropy PMFs are shown for the development and evaluation subsets. While there is still a substantial degree of overlap between the two PMFs, there is a certain shift between the two PMFs. Given the specific design goals of the RSR 2015 dataset this differences is not necessarily problematic; the database was designed such that the devices used in recording the evaluation subset are different to those used in the recording of the development subset. In this respect, differences in the two PMFs would be expected by design. The entropy validation tool could then be used in order to design a protocol with varying *difficulty*, where an indication of difficulty could be provided by the shift between the two PMFs.

### 4.3. ASVspoof 2015

Dataset entropy PMFs for the ASVspoof 2015 dataset are illustrated in Fig. 3. Two different plots are shown, both for training,

[2] https://sites.google.com/site/thereddotsproject/

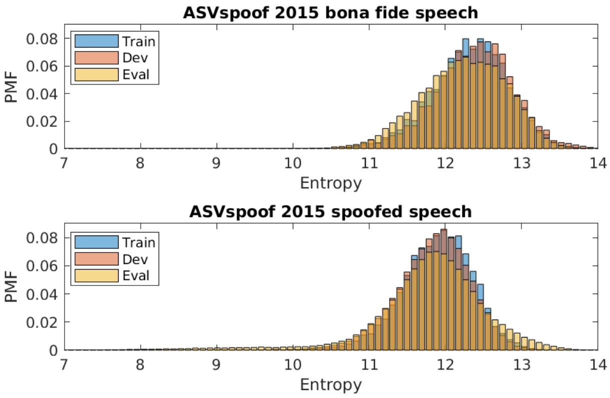

Figure 4: *As for Fig. 3 except for the ASVspoof 2015 database of bona fide speech (top) and speech synthesis and voice conversion spoofing attacks (bottom).*

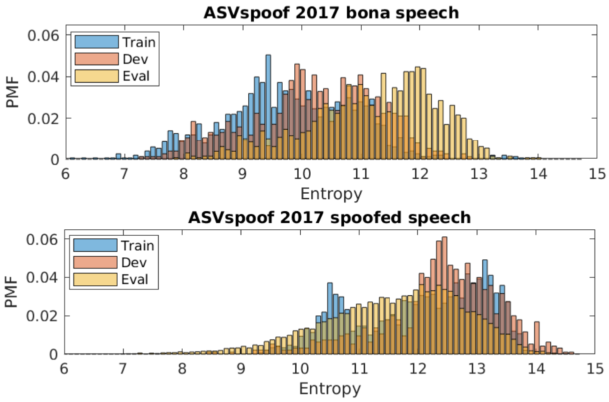

Figure 5: *As for Fig. 4 except for the ASVspoof 2017 v2.0 database of bona fide speech (top) and replay spoofing attacks (bottom).*

development and evaluation partitions. The top plot show three PMFs for bona fide data whereas the the bottom plot shows results for spoofed data. Whereas the three PMFs are more or less identical for bona fide data, differences are observed for the spoofed data. Differences would not be expected for bona fide data but, again according to design, differences would be expected for spoofed data. In order to mimic spoofing *in the wild*, the ASVspoof 2015 was designed such that speech synthesis and voice conversion spoofing attacks used to create the evaluation subset are different to those used to create the training and development subsets. These results again indicate that the protocols serve the intended purpose and the entropy validation tool could be used to judge the difficulty of the task.

### 4.4. ASVspoof 2017

Results for the ASVspoof 2017 v2.0 database are illustrated in Fig. 5 for the same data splits in Fig. 4. These results show somewhat different findings that those reported above. The three PMFs for spoofing attack data again show differences, differences which are expected by design. In contrast to those for the ASVspoof 2015 database, the PMFs for bona fide speech are different. This is an unexpected result since the definition of bona fide speech should be consistent across the three datasets. This may be a consequence of the heterogeneous nature of the data (different acoustic conditions and devices), or it may be an indication of some other external influence that is poorly marginalised. Whatever the reason, this finding would suggest that the ASVspoof 2017 data partitions are not well balanced.

## 5. Conclusions and discussion

This paper presents a simple and efficient approach to speech database and protocol validation. Based on straightforward measures of waveform entropy, the entropy validation tool reported in this paper can help to highlight potentially subtle dataset imbalances such as those between training, development and evaluation datasets.

Findings stemming from the application of the waveform entropy validation tool to four standard database show that the TIMIT database is well balanced; training and development subsets exhibit an almost identical nature.

Results for the RSR 2015 database show a subtle shift between the development and evaluation subsets. These differences are however to be expected. They are the consequence of design; in order to assess channel-robustness the two subsets were collected with different acquisition devices.

The ASVspoof 2015 database shows similar entropy distributions for different partitions of bona fide data, whereas those for spoofed data show some divergence. Once again, this is expected by design; in order to assess generalisation in the wild, evaluation data was collected using different spoofing algorithms.

Findings for the ASVspoof 2017 v2.0 database are rather different. First, they show substantial variation between spoofed data in training, development and evaluation partitions. This is once again the result of design; the ASVspoof 2017 v2.0 database was collected with highly varying acoustic conditions, e.g. background noise, using many different acquisition devices, with each combination exerting varying influences on the speech signal. It is for this reason that, unlike those for other databases, entropy distributions for the ASvspoof 2017 v2.0 database are non-Gaussian in nature; there is a lack of statistics to account properly for the variation. Second, and more interestingly, differences are also observed for bona fide speech partitions. This is not expected; there is only one definition of bona fide speech and thus the entropy distributions across training, development and evaluation should be similar.

Findings stemming from the use of such a protocol might be difficult to interpret correctly. The protocol validation tool used in this paper could be used to investigate or correct such observed imbalances, just as it could be employed to fine tune or vary the difficulty corresponding to a given database and protocol for a specific application.

To the best of our knowledge, this is the first investigation of database and protocol validation tools. While admittedly a rather naive evaluation and the first steps in this direction, the nonetheless shows the importance of and potential for thorough and adequate validation.

# 6. References

[1] I. Shafran and R. Rose, “Robust speech detection and segmentation for real-time ASR applications,” in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, Hong Kong, April 2003, pp. 432–435.

[2] J. Gemmeke, T. Virtanen, and A. Hurmalainen, “Exemplar-based sparse representations for noise robust automatic speech recognition,” *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 7, 2011.

[3] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[4] J. Joseph P. Campbell, “Speaker recognition: A tutorial,” *Proc. of the IEEE*, vol. 85, no. 9, pp. 1437 – 1462, 1997.

[5] M. Przybocki and A. Martin, “NIST speaker recognition chronicles,” in *Proc. Odyssey 2004: the Speaker and Language Recognition Workshop*, May 2004, pp. 15–22.

[6] Z. Wu, T. Kinnunen, N. Evans, and J. Yamagishi, “Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2015) Database,” 2015, http://dx.doi.org/10.7488/ds/298.

[7] T. Kinnunen, M. Sahidullah, H. Delgado, Todisco, Massimiliano, N. Evans, J. Yamagishi, and K. A. Lee, “The 2nd Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2017) Database, Version 2,” 2018, http://dx.doi.org/10.7488/ds/2313.

[8] M. Wester, Z. Wu, and J. Yamagishi, “Analysis of the voice conversion challenge 2016 evaluation results,” in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, 2016, pp. 1637–1641. [Online]. Available: https://doi.org/10.21437/Interspeech.2016-1331

[9] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. MüLler, and S. Narayanan, “Paralinguistics in speech and language-state-of-the-art and the challenge,” *Comput. Speech Lang.*, vol. 27, no. 1, pp. 4–39, Jan. 2013. [Online]. Available: http://dx.doi.org/10.1016/j.csl.2012.02.005

[10] A. K. Jain, A. Ross, and S. Prabhakar, “An introduction to biometric recognition,” *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 14, no. 1, pp. 4–20, Jan. 2004. [Online]. Available: http://dx.doi.org/10.1109/TCSVT.2003.818349

[11] P. Lieberman and S. Blumstein, *Speech Physiology, Speech Perception and Acoustic Phonetics*, ser. Cambridge studies in speech science and communimication. Cambridge University Press, 1991. [Online]. Available: https://books.google.fr/books?id=dYa4AQAACAAJ

[12] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, “Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1,” *NASA STI/Recon technical report n*, vol. 93, 1993.

[13] A. Larcher, K. A. Lee, B. Ma, and H. Li, “Text-dependent speaker verification: Classifiers, databases and rsr2015,” *Speech Communication*, vol. 60, pp. 56 – 77, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639314000156

[14] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, “ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements,” 2018, submitted. [Online]. Available: http://www.asvspoof.org/data2017/ASVspoof2017_V2_Odyssey_2018.pdf

[15] O. Ben-Harush, O. Ben-Harush, I. Lapidot, and H. Guterman, “Initialization of iterative-based speaker diarization systems for telephone conversations,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 414 –425, feb. 2012.

[16] I. Lapidot and J.-F. Bonastre, “Generalized viterbi-based models for time-series segmentation applied to speaker diarization,” in *ODYSSEY 2012 -The Speaker and Language Recognition Workshop*, 2012.