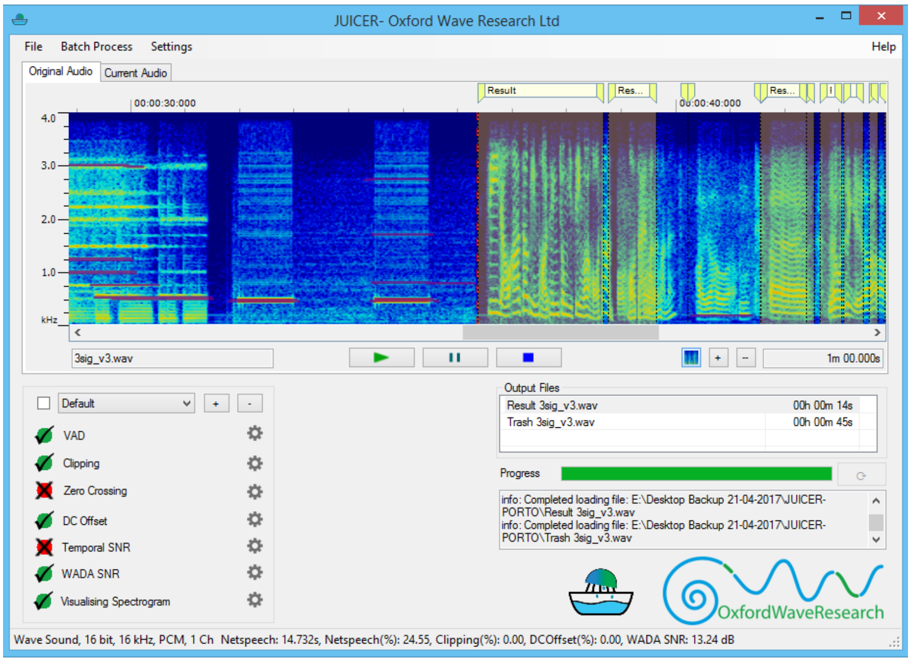
Estimating the Good, the Bad and the Ugly in Speech Recordings

Alankar Atreya, Oscar Forth, Samuel Kent, Finnian Kelly, and Anil Alexander
Research and Development, Oxford Wave Research Ltd, Oxford, United Kingdom.
{alankar|oscar|sam|finnian|anil}@oxfordwaveresearch.com

Audio recordings commonly encountered in forensic and investigative casework can vary widely in both subjective quality and relative suitability for automatic speech and speaker recognition. Qualitative extrinsic variations may be related to the file format, i.e. codecs, sampling rates, bit rates etc., or to the acoustic content within the files, such as the channel and microphone, the presence of interfering tones, background noise or music, and the net amount of speech present. In performing a forensic speaker comparison, it is important to be aware of these variations and their potential influence on the results. In this paper, we explore how objective measurements can be used to triage a dataset to inform automatic speaker recognition system performance.

VoxCeleb (Nagrani et al., 2017) is a large-scale 'in the wild' database of over a thousand celebrity speakers collected from Youtube by an Oxford University group. This relatively unconstrained database contains a variety of recording conditions and various intrusive noises, making it an ideal test-bed for analysing the effect of objective quality measures on speaker comparisons.



☐ Default

✓ VAD
☒ Zero Crossing
✓ Clipping
✓ DC Offset
☒ Temporal SNR
✓ WADA SNR
✓ Visualising Spectrogram

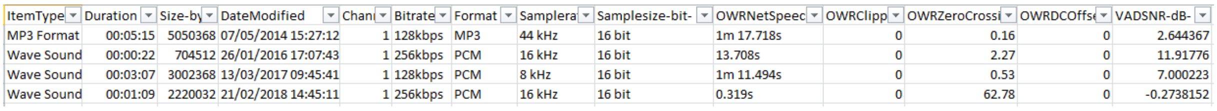
Result 3sig_v3.wav	00h 00m 14s
Trash 3sig_v3.wav	00h 00m 45s

```
info: Completed loading file: E:\Desktop Backup 21-04-2017\JUICER-
PORTO\Result 3sig_v3.wav
info: Completed loading file: E:\Desktop Backup 21-04-2017\JUICER-
PORTO\Trash 3sig_v3.wav
```

Wave Sound, 16 bit, 16 kHz, PCM, 1 Ch Netspeech: 14.732s, Netspeech(%): 24.55, Clipping(%): 0.00, DCOffset(%): 0.00, WADA SNR: 13.24 dB

Figure 1. Screenshot of JUICER software for audio quality metric extraction

For the experiments in this paper we used JUICER software (screenshot in Figure 1) for audio quality metric extraction. JUICER allows the user to quickly triage large numbers of audio files for multiple quality metrics, including voice activity, clipping, voicing, WADA-SNR (Waveform Amplitude Distribution Analysis Signal-to-Noise Ratio) (Kim and Stern, 2008), and tone detection. These metrics (a selection is shown in Figure 2) were used to group audio files according to their objective quality. VOCALISE 2017B (Alexander et al., 2016) was then used to assess speaker recognition performance across the different groups.



ItemType	Duration	Size-by	DateModified	Chan	Bitrate	Format	Samplera	Samplesize-bit-	OWRNetSpeec	OWRClipp	OWRZeroCrossi	OWRDCOffs	VADSNR-dB-
MP3 Format	00:05:15	5050368	07/05/2014 15:27:12	1	128kbps	MP3	44 kHz	16 bit	1m 17.718s	0	0.16	0	2.644367
Wave Sound	00:00:22	704512	26/01/2016 17:07:43	1	256kbps	PCM	16 kHz	16 bit	13.708s	0	2.27	0	11.91776
Wave Sound	00:03:07	3002368	13/03/2017 09:45:41	1	128kbps	PCM	8 kHz	16 bit	1m 11.494s	0	0.53	0	7.000223
Wave Sound	00:01:09	2220032	21/02/2018 14:45:11	1	256kbps	PCM	16 kHz	16 bit	0.319s	0	62.78	0	-0.2738152

Figure 2. A selection of objective quality metrics obtained from JUICER

A subset of 10800 files from 1174 speakers was selected from VoxCeleb. The files were then grouped according to the WADA-SNR and the minimum duration of net speech present within them. To select three groups of files based on their quality, we calculated the mean and standard deviation of the SNRs of all the files in the database and set the cut-off point at 1 standard deviation above and below the mean: poor files were defined as having an SNR below 11.5 dB, medium files as having an SNR between 11.5 dB and 23.5 dB and good as having an SNR above 23.5 dB. As greater than 70% of files fell into the medium group, a subset of the medium group was randomly selected to provide a more balanced comparison of the 3 groups.

Table 1. Convex Hull Equal Error Rate (EER), log-likelihood ratio cost (Cllr), and minimum log-likelihood ratio cost (Cllr-min), for three WADA-SNR-defined groupings of the VoxCeleb Corpus. Cross-validation score calibration was applied using BioMetrics software.

Group	*No. of speakers*	*No. of files*	*EER%*	*Cllr*	*Cllr-min*
Poor	720	1322	8.63	0.298	0.273
Medium	772	1330	5.10	0.202	0.177
Good	669	1300	3.78	0.139	0.126

In Table 1, error metrics for each of the three groupings is provided. There is a clear relationship between objective quality in terms of WADA-SNR and speaker recognition performance. This serves as an example of using quality metrics as a gatekeeper to an automatic speaker recognition system. In addition to this *per-file* gatekeeping role of quality metrics, we have observed benefits from using *within-file* quality metrics to select portions of files that satisfy various objective criteria, particularly in the case of long surveillance-like files that contain relatively little speech.

References

Alexander, A., Forth, O., Atreya, A. A., and Kelly, F. (2016). VOCALISE: A forensic automatic speaker recognition system supporting spectral, phonetic, and user-provided features. *Speaker Odyssey 2016*, Bilbao, Spain.

Kim, C., and Stern, R. M. (2008). Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. *Interspeech 2008*, Brisbane, Australia.

Nagrani, A., Chung, J. S., and Zisserman, A. (2017). VoxCeleb: a large-scale speaker identification dataset. *Interspeech 2017*, Stockholm, Sweden