DOI: 10.17469/O2108AISV000003

KATHARINA KLUG, MICHAEL JESSEN, YOSEF A. SOLEWICZ, ISOLDE WAGNER

Collection and analysis of multi-condition audio recordings for forensic automatic speaker recognition

The major aim of the project presented here is to compile a corpus from real case recordings to validate more recording conditions and languages under match and mismatch conditions for forensic automatic speaker recognition (FASR). The challenges and limitations of compiling a real case corpus are explained. First results of validation tests are presented for male speakers of German in the match condition [voice message – voice message] as well as in the mismatch condition [voice message – telephone]. Results for the match condition [voice message] are compared to previous findings for the match condition [telephone]. Variations of performance metrics such as Equal Error Rate (EER) and log-likelihood-ratio cost (Cllr) are discussed with respect to effects of normalisation and calibration, and patterns of score distributions are analysed using Tippett plots.

Keywords: Forensic automatic speaker recognition, Real case recordings, Validation, Match and mismatch condition, Calibration.

1. *Introduction*

The present paper outlines a research project aiming to validate forensic automatic speaker recognition (FASR) systems by using real forensic audio recordings. With the improvement of FASR systems over the past two decades, the use of the automatic approach as an element alongside auditory and/or acoustic analysis in speaker comparison casework has increased significantly worldwide (Gold, French, 2019). Both the auditory-acoustic approach and FASR are based on Bayesian principles in assessing similarity and typicality (Jessen 2018; Rose 2002 for the foundations of Bayesian statistics and the likelihood ratio in forensic speaker comparison). However, while FASR generates a quantitative strength-of-evidence outcome (likelihood ratio), the auditory-acoustic approach produces a more qualitative outcome. Combining both approaches would therefore increase the level of information when expressing conclusions.

Validating (i.e. testing) FASR systems is necessary to assess the case-specific strength of evidence given the specific case condition material (Drygajlo, Jessen, Gfroerer, Wagner, Vermeulen & Niemi, 2015: 17; Morrison, Enzinger, Hughes, Jessen, Meuwly, Neumann, Planting, Thompson, van der Vloed, Ypma, Zhang, Anonymous & Anonymous, 2021). So far, only a limited proportion of common forensic scenarios in cases submitted to the German Federal Criminal Police Office

(Bundeskriminalamt, BKA) has been validated, i.e. [telephone][1] conversations of male speakers of German. Accordingly, the rate of FASR application in casework is low. Within the last 15 years only 10-20% of conducted speaker comparison cases at the BKA have met the requirements for FASR application. These requirements include sufficient recording quality and quantity of speech, and consistency in terms of spoken language and recording condition. Over the last five years, the percentage of FASR-treated speaker comparison cases actually decreased. Possible reasons for the reduction could be the change in submitted recording material from predominantly [telephone] recordings to different types of recordings (e.g. voice messages, videos, interior surveillance). The vast majority of scenarios, i.e. languages and recording conditions, and in particular the various combinations of mismatch situations, have not been tested yet. The present research project starts to fill this gap by compiling a corpus consisting of speakers of various languages and recording conditions in match and mismatch settings.

However, given the vast number of potential combinations of mismatch scenarios that forensic speech experts might face, the present project is merely a starting point. The collection of real forensic audio recordings needs to be continuously expanded to fully exploit the capability of such a real case corpus. Due to restrictions in data availability and data access this remains the main challenge for validating FASR on real forensic audio recordings.

2. *Mismatch conditions and unfamiliar match conditions*

Mismatch factors can be manifold. Kelly, Hansen (2021) subdivide (1) extrinsic and (2) intrinsic factors, which are either (1) speaker-independent (e.g. background noise, transcoding effects, channel characteristics, recording devices) or (2) speaker-dependent (e.g. emotional or stylistic variation, vocal effort, short- or long-term voice fluctuations as in the case of a cold or ageing effects). Due to this diversity, more than one mismatch factor can appear at the same time, leading to 'multi-mismatch' scenarios. In fact, multi-mismatch is inevitable in forensic speaker comparison cases and, furthermore, the ground truth of many mismatch factors, such as transcoding history, recording hardware and software, date of origin, can only be estimated.

Previous FASR studies investigated the effects of individual types of mismatch using controlled and/or manipulated data (e.g. Nash, 2019; Kelly, Hansen, 2021). Such simulated test corpora are necessary to arrive at generalisations as the data's ground truth is known and potential further factors can be kept as similar as possible to investigate only one factor at a time. Factors which have been studied so far include but are not limited to: mismatch in spoken language (Künzel, 2013), mismatch in time span between recordings of interest (Morrison, Kelly, 2019), mismatch in acoustics (net speech duration, signal-to-noise ratio, reverberation,

[1] In this paper, the conditions tested using FASR are listed in square brackets.

frequency bandwidth and transcoding) (Nash, 2019), mismatch in recording device (van der Vloed, Kelly & Alexander, 2020) and mismatch in vocal effort (Kelly, Hansen, 2021).

Nash (2019) studied the effect of net duration on discrimination performance and found that lower net duration led to lower same-speaker (subsequently abbreviated as SS) scores and higher different-speaker (DS) scores, while system accuracy decreased (p. 126). Testing the effect of net duration in match and mismatch conditions using an i-vector/PLDA system he found that down to a performance tipping point of 10 seconds the discrimination performance decreased gradually. Net durations below 10 seconds, however, resulted in an exponential performance degradation. Mismatch in net duration does not seem to have an impact on the Equal Error Rate (EER), i.e. discrimination performance was not systematically improved by matching the net duration of the recordings to be compared. According to Nash's results, EERs increased when net durations decreased from 25s, over 20s to 10s, even though they matched in duration for each of these steps. But, for example, having a pair of recordings with one of them 10s long and the other 30s (duration mismatch) led to better average performance than matching both recordings at a value of 10s. Duration might be the only variable that behaves in this manner. Usually, a mismatch in conditions led to poorer performance than a match. For example, a mismatch between noise-free and noise-degraded speech is worse than a match between two equally noise-degraded recordings.

Beyond mismatch, a further issue is the behaviour of "unfamiliar" match conditions. The classifier "unfamiliar" refers to types of intrinsic factors, extrinsic factors, or combinations of the two which are scarcely represented in the literature and have not been sufficiently tested yet (at the BKA, Israel Police or other forensic services). In contrast to mismatch, here conditions are matched but unfamiliar (i.e. same but unfamiliar condition for the questioned speaker as well as for the suspected speaker). Forensically, the question arises whether a test (validation) that has been conducted for Condition A also applies to Condition B, i.e. whether the results for the two conditions are very similar (e.g. Solewicz, Jessen & van der Vloed, 2017 for strong similarity between a Dutch-based and a German-based test), or whether at least some process of interpolation or extrapolation can be applied in order to bridge the two (Morrison, Kelly, 2019 for an example of interpolation).

To give an example of unfamiliar match conditions, Nash (2019) investigated the effect of transcoding using controlled data of twelve different codec types with various settings. Lossy codecs in particular, relying on psychoacoustic compression, influenced the discrimination rate considerably and increased the rate of false rejections and false acceptances (p. 264). Formats typically involved are MP3, M4A and AAC. Nash hypothesised that recordings from the same codec would artificially increase discrimination performance due to a match in channel characteristics (p. 256). This means that a match among unfamiliar

conditions might lead to overconfident results – a pattern that will become relevant in this study.

A recent study on both mismatch and unfamiliar match conditions was carried out by Kelly, Hansen (2021). They studied the impact of vocal-effort variation using [whisper], [Lombard] speech and [neutral] speech to assess mismatch in vocal effort. They found, using an x-vector system, that both mismatch conditions [whisper – neutral] and [Lombard – neutral] lowered discrimination performance; however, more so when [whisper] was involved (EER of ≈ 20%) rather than [Lombard] speech (EER of max 3.62%). Considerably better discrimination performance was found for [non-neutral] match comparisons, i.e. [whisper – whisper] (EER of max 7.36%) and [Lombard – Lombard] (EER of max 2.12%). Thus, the match conditions in vocal effort – even though they are based on [non-neutral] speech – outperformed the mismatch situation.

The above studies allow the effects of individual factors to be assessed under controlled conditions. However, given that controlled data produce very homogeneous scenarios, the results obtained should be considered as more optimistic than expected for typical forensic scenarios in which factors of interest do not occur in isolation. Therefore, as Drygajlo et al. (2015: 17) demanded, FASR systems need to be validated using corpora that match casework conditions, if possible, by using available data from previous casework. In the casework-based paper presented here, the mismatch condition was [voice message] against [telephone] communication and the "unfamiliar" match condition was [voice message]. For the mismatch condition it is interesting to see how performance changed (probably decreased) relative to a [voice message] match condition. For the match condition based on [voice message] data, a comparison was made to a previously used [telephone] interception corpus (Solewicz et al. 2017) and the question was whether performance is similar in these two match condition datasets. "Performance" not only concerns speaker discrimination or calibration values, but also the actual distributions of the scores in Tippett plots, which are expected to reveal further interesting patterns.

3. *The FORBIS project*

The FORBIS project (short for *Collection and analysis of recordings for FORensic BIometric Speaker recognition*) is a research project funded by the EU Internal Security Fund and was conducted at the BKA from March 2019 to August 2021. The project's main goal was to compile a corpus of recordings suitable for FASR validation tests to enable the application of FASR for additional recording conditions and further spoken languages beyond the hitherto focus on German telephone data. The basis for the compilation of the corpus was the archive of the BKA section 'Text, speech and audio', from which suitable recordings were collected. In addition, other BKA divisions were asked to provide real forensic audio and video material from investigative proceedings.

3.1 Corpus limitations

Potential extrinsic and intrinsic factors, causing multi-mismatch, should be considered when compiling a corpus based on real case recordings. The following list shows some examples of relevant factors (for a more complete collection see Hansen, Bořil, 2018).

Extrinsic factors:

- Recording quantity: net speech duration
- Recording quality: effects of recording and transmission procedures, i.e. linear and non-linear distortions
- Background noise, e.g. traffic, babble, wind noise

Intrinsic factors:

- Emotional state
- Speaking style
- Vocal effort
- Time span between recordings

Because the actual basis, i.e. the ground truth, of many factors remained unknown (e.g. how the recording was transmitted exactly, or if a speaker was indeed emotional), evaluating the impact of individual factors was complicated. Further difficulties were caused by various imbalances, namely the number of recordings per speaker, the number of speakers per recording condition (e.g. telephone intercept, voice message) and the number of speakers per language. As for any collection of forensically authentic recordings, there was also a remaining uncertainty about speaker identities, i.e. correct assignment of SS and DS status in the validation dataset.

3.2 Corpus compilation

Only recordings of male speakers were collected. In addition to German, the languages Arabic, Turkish and Russian most frequently appeared in casework at the 'Text, speech and audio' section of the BKA in recent years and thus were included in the corpus. Typically, the following four recording conditions were submitted for speaker comparison purposes and thus required validation tests: [telephone] interception, [voice message], [video] and [interior surveillance] (both indoor and car surveillance). To prevent the dominance of a few speakers, the maximum number of recordings per speaker was limited to six recordings per language and condition. In general, the aim for the corpus was to arrive at a minimum of 20 speakers for each language and recording condition. In addition to the recordings collected for creating a test set, matching reference populations were needed for normalisation purposes within the FASR system used (addressed in § 4). The aim was to collect 30 additional speakers with one recording per speaker for each language and each recording condition.

4. *FASR system and performance testing*

The FASR system VOCALISE (VOice Comparison and Analysis of the LIkelihood of Speech Evidence) Version 2.7 (Kelly, Forth, Kent, Gerlach & Alexander, 2019a for further information) was used to conduct the speaker recognition tasks (i.e. speaker comparisons) based on Mel Frequency Cepstral Coefficients (MFCCs). Pre-tests showed that the state-of-the-art x-vector PLDA system based on DNN embeddings outperformed the previous i-vector PLDA system. According to recommendations by the developers x-vectors were supposed to improve system performance under mismatch conditions and when degraded recordings were used. The results presented here were taken from the x-vector system only. Bio-Metrics Version 1.8 was used for examination of the system's discrimination and calibration performance (Equal Error Rate, EER, and log-likelihood-ratio cost, Cllr), visualisation of the score distributions (Tippett plots), implementation of cross validation calibration (CV) and Zoo plots.

Various system options were tested individually as well as in combination, revealing the system's raw scores [-norm, -CV], normalised scores [+norm, -CV] as well as calibrated scores for both, raw scores [-norm, +CV] and normalised scores [+norm, +CV]. "Score Normalisation" (henceforth "normalisation") was applied using S-norm (symmetric normalisation, Shum, Dehak, Dehak & Glass, 2010). The details of the normalisation procedure in VOCALISE are described in Kelly et al. (2019a). Essentially, the recordings used in the test were compared with recordings of a set of speakers unrelated to the test data but with equivalent speaking conditions. Means and standard deviations of the score results of these comparisons were calculated and used to shift the score result of each comparison in the test. Generally, normalisation has the effect of improving discrimination performance to some extent, which is the main motivation for its use.

Calibration was applied using logistic regression cross validation calibration. Calibration has the goal of turning the raw scores into interpretable likelihood ratios, which means that if expressed in terms of LLR (log likelihood ratios) values larger than zero are typical for SS comparisons and values smaller than zero are typical for DS comparisons. The logistic regression calibration technique calculates shift and scale parameters from a calibration dataset and applies them to the scores. Cross validation is a "leave-one-out" method. The same corpus is used both for training the calibration parameters and for testing, but the speaker(s) involved in each comparison cycle is/are removed from the training set. For further literature on EER, Cllr, Tippett plots, logistic regression calibration and the cross validation principle see van Leeuwen, Brümmer (2007), Morrison (2011), Morrison (2013) and Drygajlo et al. (2015).

5. *Speech data*

Validation tests were performed on recordings of German speakers in the match condition [voice message], as well as in the mismatch condition [voice message –

telephone] interception. Available material gained from the FORBIS corpus was fully exploited. The material used for the tests is summarised in Tab. 1. The target criteria for the number of speakers could be met for the match condition (20 speakers), but not for the mismatch condition. Only recordings of 7 speakers could be collected in both conditions: [voice message] and [telephone] interception. For the reference population, recordings of 15 additional speakers from the [voice message] condition could be collected. For the mismatch condition, recordings of another 15 speakers from [telephone] interceptions were used. Consequently, the reference population for the mismatch condition was balanced for both conditions, as is the recommended procedure when using S-norm under mismatch conditions.

5.1 Pre-tests

Transcoding formats based on psychoacoustic compression (e.g. MP3, AAC) are known to degrade discrimination performance using controlled data (Nash, 2019). Pre-tests confirmed these results. Recordings transcoded in MP3 or AAC format produced some outliers. The OPUS format, also available in high quality, did not show such an effect. However, this impression is based on only 23 OPUS files in the match corpus and 6 OPUS files in the mismatch corpus.

Originally, the corpus creation aimed at maximising the number of speakers and the number of recordings per speaker. Therefore, a rather permissive selection process was applied in terms of suitable recording formats, recording quality and neutral speaking style. Accordingly, the system performance in terms of EER of the original test set was quite poor.

The following types of recording degradations were identified in the pre-tests as causing lower system performance and were therefore removed from the test set:

- Reduced frequency bandwidth
- Unsteady sound pressure level, i.e. clipping and reduction
- Data loss due to transcoding
- Non-neutral speaking style, e.g. emotional, fatigue, intoxicated

The exclusion of recordings was based on auditory and acoustic examination. Despite excluding poor and non-neutral recordings, the compiled test set (Tab. 1) still contained different kinds and degrees of degradation causing random effects of mismatch. This, however, is typical of casework data.

For data used in the mismatch test, a slight tendency was observed that there was a greater proportion of quality degradation in the [voice message] recordings than in the [telephone] recordings. The quality degradations were in-vehicle noise, background speakers and reverberation, all of which were either absent or present to a lower degree in the [telephone] recordings. This tendency will become relevant when interpreting the results.

The recordings compiled for the test sets also underwent a screening process using Zoo plots (see Dunstone, Yager, 2009). In the data set intended for the mismatch test, one outlier in the form of a "phantom" was observed. Phantoms are

speakers that show an unusually low similarity (in terms of LLRs) when compared to other speakers and also when compared to themselves, i.e. across recording of the same speaker. That phantom-type speaker was the only one among the seven who was classified as non-normal in the Zoo plots; this status stayed the same across the four conditions [-norm, -CV], [-norm, +CV], [+norm, -CV] and [+norm, +CV]. Rather than excluding that speaker from analysis altogether, two separate types of tests were performed, one in which that speaker (referred to as "P") was included (see third column in Tab. 1) and one in which he was excluded (see last column). This seemed advisable given the number of available speakers was quite small from the start.

Table 1 - *Test sets for match and mismatch condition*

Number of	Match	Mismatch	Mismatch without speaker P
Speakers	20	7	6
Recordings	62	34	29
SS comparisons	42	39	33
DS comparisons	798	246	175
Ref. population speakers	15 (all voice messages)	30 (balanced for both rec. condition)	30 (balanced for both rec. condition)

6. *Methodology*

Each recording was converted into a PCM WAV format (44.1 kHz, stereo) using an in-house audio converter. All files were downsampled to 8 kHz using Praat (Boersma, Weenink, 2016) as only frequencies up to 4 kHz are used for MFCC extraction (Kelly, Fröhlich, Dellwo, Forth, Kent & Alexander, 2019b). The channels with the speakers of interest were extracted and the relevant net speech was manually labelled and extracted using the TextGrid and script function within Praat. The main criterion for speech labelling was based on intelligibility. Only neutral speech was labelled while interferences and extreme occurrences of non-neutral speech were discarded, e.g. disturbing noise, further speakers in one-channel recordings, non-neutral phonation types (falsetto, whisper), highly increased vocal effort (screaming etc.), laughing. Net speech editing included the manual removal of pauses; VAD (voice activity detection) was therefore disabled in VOCALISE for the purpose of the tests. Nash (2019: 259) observed a cliff edge effect for recordings with net duration lower than 10 seconds. Especially [voice message] recordings are often characterised by limited amounts of net speech. Therefore, the minimum required net speech duration was set to 10 seconds, while a maximum of 60 seconds was chosen.

Personal data (e.g. names, addresses, telephone numbers, specific places) were removed. Intrinsic and extrinsic factors were documented, such as vocal effort,

technical degradations (reduced frequency bandwidth, data loss in mid-frequency regions, estimated microphone distance, reverberation, clipping, distortion) and background noise (babble, wind, traffic). Recordings with background music, often present when speech is recorded in bugged cars, were discarded. Additionally, it was indicated whether the degradation was temporary or permanent during the recording, and whether the recording qualified for investigation of language mismatch and/or recording condition mismatch because the respective speaker was also recorded in another spoken language and/or recording condition.

7. *Results and discussion*

Results are shown in the form of Tippett plots that represent the score distributions of the tests (Figs. 1-6). Tab. 2 contains information about the numerical performance parameters EER and Cllr.

Within the software Bio-Metrics, that was used to generate the Tippett plots, the representations are called Equal Error plots. For current purposes this is equivalent to Tippett plots. In a Tippett plot the line rising from left to right represents the SS results, the one falling from left to right represents the DS results. The values on the x-axes of the plots are given as log likelihood ratio (LLR, expressed in terms of natural logarithm). Without calibration [-CV] (Figs. 1, 2, 5 and 6), the values can be interpreted as LLRs in a technical sense; only after calibration [+CV] (Figs. 3 and 4) they are LLRs in a forensically literal sense (where LLR = zero indicates maximally neutral evidence). The y-axes of the Tippett plots show error rates. The point where the SS and the DS lines intersect is known as the Equal Error Rate (EER). The higher the intersection on the level of the y-axis (hence the higher the EER), the worse the speaker-distinguishing ability of the respective system (the technical term for this is speaker discrimination). All upcoming Tippett plots representing the mismatch condition refer to the test set that included speaker "P" (the phantom speaker in the Zoo plots). The impact of excluding this speaker on the Tippett plots are not visualised in Figs. 1 and 2 but are verbalised (this does not apply to the calibrated results in Figs. 3 and 4, which show analogous effects). However, the numerical impact is fully documented in Tab. 2.

Figure 1 - *Tippett plot: raw scores [-norm, -CV] in the match condition [voice message] (solid green) and the mismatch condition [voice message – telephone] (dotted pink)*

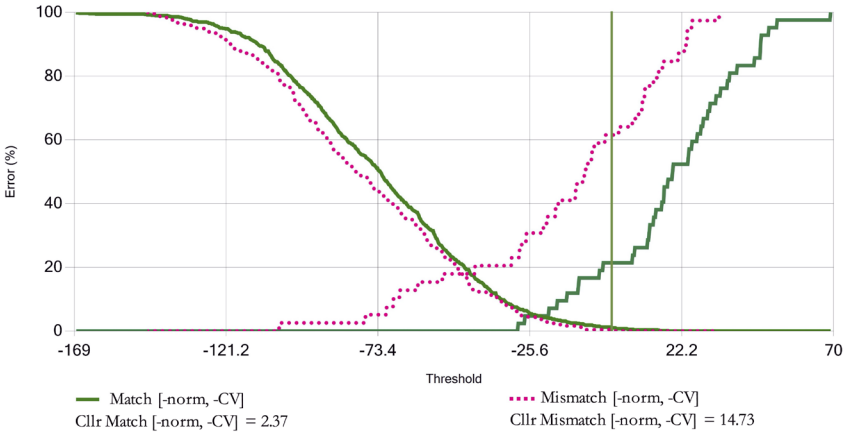


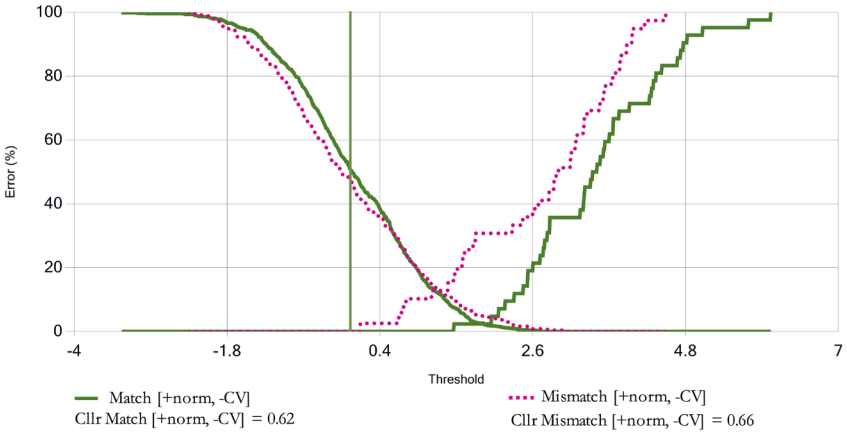
Fig. 1 shows the results of the raw scores [-norm, -CV]. The solid green lines represent the [voice message] comparisons in the match condition. The dotted pink lines represent the comparisons in the mismatch condition [voice message – telephone]. The green vertical line shows the location on the x-axis where LLR = 0 (applying to this and all subsequent figures). When match and mismatch conditions are compared in Fig. 1, there is very little difference in the DS distribution but a strong difference in the SS distribution. In the DS distribution the values (referred to as scores) are slightly smaller under the mismatch than under the match condition. In the SS distribution the scores are much smaller under the mismatch than under the match condition. Since the scores of the mismatch condition (compared to the match condition) move leftwards in the SS distribution but stay fairly constant in the DS distribution, the degree of overlap of the distributions is increased. Consequently, the intersection point of the distributions moves upwards on the y-axis and therefore the EER of the mismatch condition increases, i.e. speaker discrimination deteriorates.

Essentially, the same pattern is shown in Fig. 2, which illustrates the equivalent results after applying normalisation with S-norm, but without calibration [+norm, -CV]. Here, any difference among the DS scores has practically reduced to zero, but there is still a clear difference among the SS scores of the same kind (lower scores under the mismatch compared to the match condition) as in Fig. 1 [-norm, -CV].

The mismatch condition results shown in Figs. 1 and 2 apply to the test in which speaker P (the Zoo plot phantom-type speaker) was included. When excluded, the mismatch SS distribution occurs closer to the SS distribution of the match test, i.e. the separation between the SS distribution of match and mismatch is reduced. On average across the score distribution, the reduction is 32 percent for raw scores and 38 percent for normalised scores. The score reduction is stronger for lower scores

than for higher scores, which also explains the striking reduction in EER due to the exclusion of P, shown in Tab. 2.

Figure 2 - *Tippett plot: normalised scores [+norm, -CV] in the match condition [voice message] (solid green) and the mismatch condition [voice message – telephone] (dotted pink)*



Previous studies investigating mismatch also found varying degrees of shifts in SS and DS scores. Mismatch regarding the time-interval between the recordings of interest was investigated by Kelly, Hansen (2016). Their score distribution plots (2016: 418, Fig. 3a, b) showed a decrease of SS scores with an increase of age difference (used age intervals: 1-10 years, 11-20 years, and 21-30 years), but almost no change among the DS scores. Kelly, Hansen (2021) studied the effect of mismatch in vocal effort based on [neutral – Lombard] and [neutral – whisper]. Again, a left-shift pattern on the SS side only was shown clearly for [neutral – Lombard], i.e. comparing match [neutral – neutral] with mismatch [neutral – Lombard] (2021: 935, Fig. 2). A possible explanation for this asymmetric behaviour could be that the effect of mismatch itself might be 'absorbed' by the dissimilarity of the DS component. This means that the difference between speakers is large enough so that a condition mismatch on top of the speaker differences is of no further consequence.

Some experiments on mismatch did not show this pattern but showed a clear lowering of scores for both SS and DS scores. Fröhlich (2017) compared recording mismatch between [telephone] interception and a mock police [interview] scenario using the forensic_eval_01 dataset (Morrison, Enzinger, 2016) with match conditions within the scenarios. Exploring her score distribution plots (2017: 72), the presented raw scores [-norm, -CV] did show a substantial left-shift for both the SS and the DS distributions in the mismatch condition compared to the match condition. The same kind of pattern can be seen for [neutral – whisper], i.e. comparing match [neutral – neutral] with mismatch [neutral – whisper] (Kelly, Hansen, 2021: 935, Fig. 2). A possible explanation for this discrepancy with the

asymmetric pattern shown above could be as follows. When the channel conditions are strong and systematic (as in forensic_eval_01) they might – simply spoken – act as a "noise-vector" that is almost uncorrelated with the "speaker-vector". For that reason, the channel influence and the (different-)speaker influence are additive. In a broader sense, [whisper] might also be conceived of as a channel effect. Like a noisy channel, it is corrupting parts of regular speech. This mode of corruption might be fairly independent of the speaker differences, and again, creating an additive effect in reducing DS scores. In contrast to this pattern, the difference between [telephone] and [voice message] investigated here does not seem to be strong, systematic, and independent enough to have a clear effect on DS scores.

Figure 3 - *Tippett plot: normalised and calibrated scores using cross validation calibration [+norm, +CV] in the match condition [voice message] (solid green) and the mismatch condition [voice message – telephone] (dotted pink)*

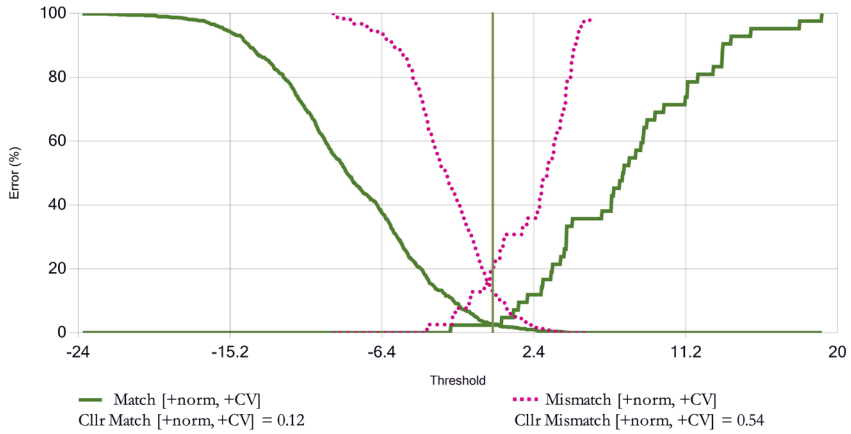


Fig. 3 is based on normalised scores followed by cross validation calibration using Bio-Metrics [+norm, +CV]. One expected effect of calibration that can be seen in the Tippett plot is that the intersection between the SS and DS distribution is located very closely at LLR = 0. Moreover, it can be observed that there is a shift towards the centre of the plot for both mismatch distributions (SS and DS). This behaviour has the effect that the values from the mismatch condition (now fully calibrated LLRs) fall entirely within the range of values from the match condition (a similar pattern can also be found in the age mismatch study by Kelly, Hansen, 2016: 2418, Figs. 3c, d). This means that under mismatch conditions it is not possible to obtain the same range of LLRs as under match conditions. Furthermore, it can be seen that the intersection point is higher on the y-axis under mismatch than match, i.e. the EER is higher.

Figure 4 - *Tippett plot: normalisation effect [+norm, +CV] (solid lines) in comparison to raw scores [-norm, +CV] (dotted lines) in the match condition [voice message] (green) and the mismatch condition [voice message – telephone] (pink)*

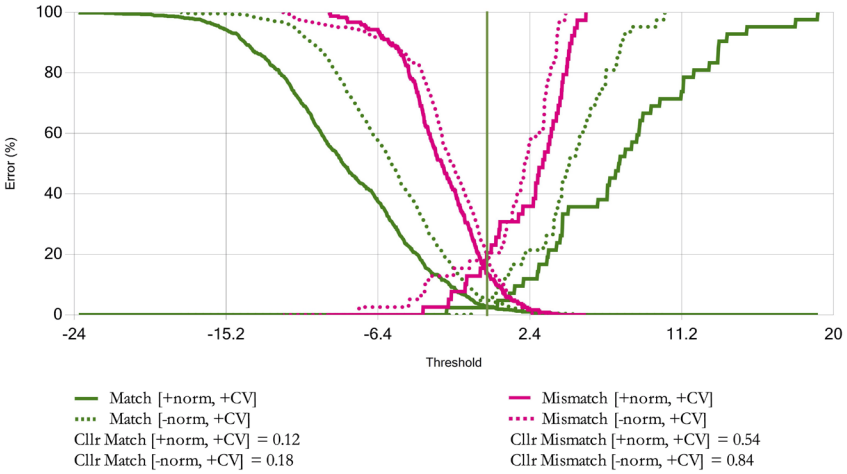


Fig. 4 shows the effect of normalisation, hence the improvement of the system's performance, on both match (green) and mismatch condition (pink). In comparison with the non-normalised raw scores (dotted lines) [-norm, +CV], the normalised distributions (solid lines) [+norm, +CV] are stretched across a wider range. This effect is stronger in the match condition (green), as the normalised scores are constantly higher for the SS and constantly lower for the DS distribution. For mismatch, the pattern is less clear especially in ranges lower than -6.4, where normalised and non-normalised results cross.

Table 2 - *Summary of EER and Cllr results for match and mismatch conditions under different system options*

Condition	System options	EER in %	Cllr	EER in % w/o speaker P	Cllr w/o speaker P
match	[-norm, -CV]	4.3	2.37	-	-
	[-norm, +CV]	4.4	0.18	-	-
	[+norm, -CV]	2.1	0.62	-	-
	[+norm, +CV]	2.2	0.12	-	-
mismatch	[-norm, -CV]	16.3	14.73	7.9	7.7
	[-norm, +CV]	17.7	0.84	10.9	0.43
	[+norm, -CV]	12.1	0.66	9.0	0.70
	[+norm, +CV]	14.9	0.54	11.3	0.41

Tab. 2 shows EERs and Cllrs for all performed tests. Within both match and mismatch condition, normalisation [+norm, -CV] has the effect of improving discrimination (lowered EER) compared to the raw scores [-norm, -CV]. This

is in line with expectation. An exception is the mismatch test in which speaker P is excluded; in this case, after normalisation, the EER is slightly higher instead of lower. Normalisation also has a certain calibration effect, i.e. unacceptably high Cllr values among raw scores are lowered to more acceptable values (smaller than 1) after normalisation. Only after applying calibration [+CV] the Cllr values are at their lowest. With sufficiently large calibration sets (the VOCALISE manual recommends at least 15 speakers with two recordings per speaker, but preferably more speakers) calibration should have a negligible effect on discrimination, i.e. EER before and after calibration should be approximately the same. This is true for the results of the match condition, but there are exceptions for the results of the mismatch condition. In the mismatch condition, the minimum recommendation for the use of calibration could not be met, which probably leads to the exceptional patterns. It is also possible that for calibration sets below the recommended minimum, the effect of calibration may be less effective because the calibration parameters may be influenced by the specific individual speaker patterns in the calibration set, rather than just reflecting the case conditions. Hence, Cllr may now be higher than if the recommendations were met.

Comparing the results of the mismatch condition with and without speaker P, it can be observed that the exclusion of this speaker causes a substantial improvement in discrimination, i.e. a reduction in the EER. Cllr is also substantially improved for the raw scores, but for the normalised scores the effects are small, or Cllr even deteriorates [+norm, -CV].

After having shown the results for mismatch compared to match, the following Tippett plots compare the unfamiliar match condition [voice message] with the more established match condition [telephone] interception. Figs. 5 and 6 compare two tests with match conditions, namely [voice message – voice message] shown earlier (represented in solid green) with [telephone – telephone] (represented in dashed blue). The [telephone] test is based on the corpus GFS 2.0 (Solewicz et al., 2017). Fig. 5 shows the results for raw scores [-norm, -CV], Fig. 6 for normalised scores (without applying calibration) [+norm, -CV].

Fig. 5 shows that on the level of raw scores there is a shift towards higher values from the condition [telephone – telephone] to [voice message – voice message]. That shift applies to both SS and DS comparison results to about the same extent. Given the complexity of an *x*-vector system, it is difficult to explain this shift. It is possible that [voice message] data are only scarcely represented in the training data of the system, whereas [telephone] conversation data are represented abundantly. As a result, it is possible that above-average similarities are found between [voice message – voice message] comparisons, resulting in higher scores. It could also be relevant that [voice message] data contains spectral information up to 7 kHz, while [telephone] data is much more limited. Accordingly, the entire 4 kHz range can be explored in [voice message] data after down sampling, which might lead to a higher score. Finally, the trends of quality degradation in the tested [voice message]

data mentioned in § 5.1 (especially in-vehicle noise) might have led to a similarity pattern that is not present in the [telephone] data.

Figure 5 - *Tippett plot: raw scores [-norm, -CV] in the match condition [voice message – voice message] (solid green) and the match condition [telephone – telephone] (dashed blue)*

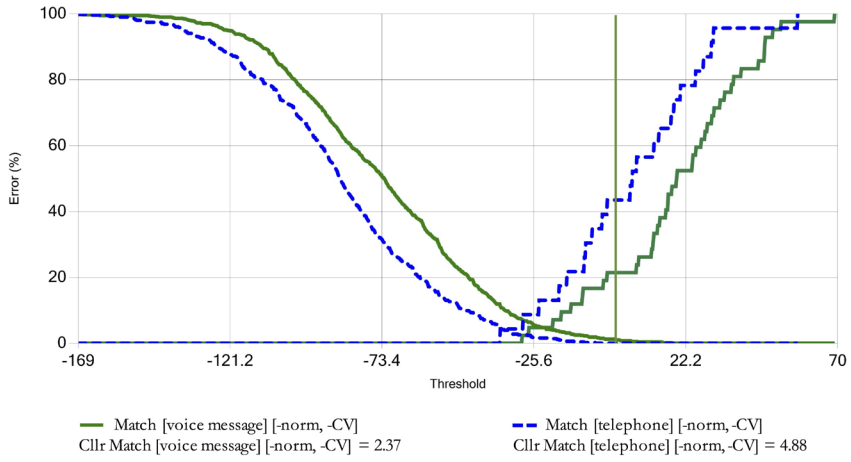


Figure 6 - *Tippett plot: normalised scores [+norm, -CV] in the match condition [voice message – voice message] (solid green) and the match condition [telephone – telephone] (dashed blue)*

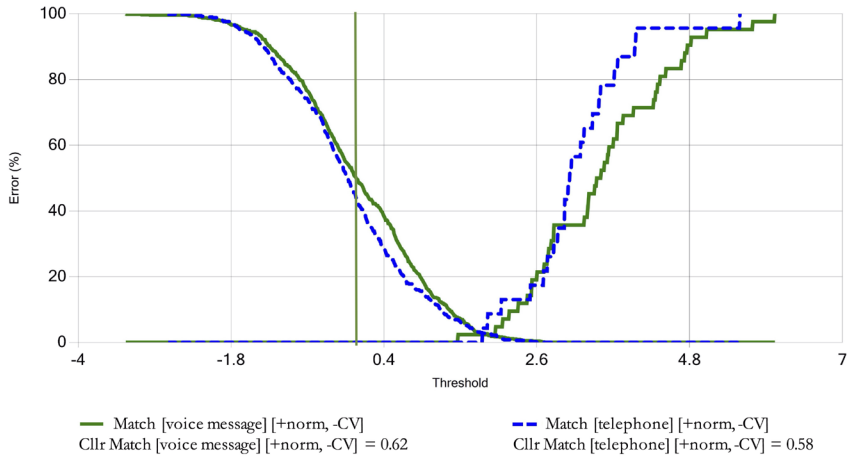


Fig. 6 shows that when normalisation is applied, the difference between the two match conditions [voice message] and [telephone] is reduced. On the DS side, the difference has almost disappeared. On the SS side, the difference is limited to only some areas of the cumulative distribution. As explained earlier, normalisation

is performed by feeding additional data into the system that corresponds to the conditions of the case. In comparisons within [voice message] data this normalisation set consists of [voice message] data, in the case of [telephone]-based comparisons, the normalisation set consists of [telephone] recordings. Providing case-relevant normalisation data seems to reduce quite strongly the differences of the score distributions that are found when no normalisation is applied, leading to a large degree of overlap between the two score distributions. About the same amount of overlap is found if cross validation calibration is applied to the normalised scores [+norm, +CV] separately for each condition (not shown here).

When looking at the similarity of the distributions in Fig. 6, one may wonder why there is such a clear mismatch effect at all. In other words, when a test based on [voice message] data leads to very similar score distributions as a test based on [telephone] interception, why is performance degraded in comparisons where the questioned speaker's recording is a [voice message] while the suspected speaker's recording is a [telephone] intercept? For an explanation one should probably consider the raw values, not the normalised scores. The raw scores are the values directly resulting from the automatic processing that takes place in the multidimensional space of the acoustic features and their models. If, as assumed above, the condition [voice message] was poorly represented in the training data of the system, these recordings might have stuck out in the multidimensional space, making them not only similar to other [voice message] recordings (as assumed above), but also dissimilar to recordings in other conditions, such as [telephone] recordings. Hence the score reductions under mismatch. The above-mentioned degradation in quality of [voice message] data compared to [telephone] data may have increased this effect.

8. *Conclusions*

The goal of the presented FORBIS project was to compile a real case forensic corpus to validate the performance of FASR under forensically realistic conditions and subsequently apply it to a greater extent in speaker comparison casework. Here, the match condition [voice message] as well as the mismatch condition [voice message – telephone] were validated. Results from the match condition [voice message] were compared to the match condition [telephone], a condition which has been far more extensively studied in casework application as well as in research projects.

When examining the results in § 7 with regard to their impact on forensic casework, the following aspects can be emphasized.

Comparison of the match condition [voice message – voice message] with the mismatch condition [voice message – telephone] has shown that when the raw scores or the normalised scores are examined, there is a reduction of SS scores from the match to the mismatch condition and essentially no change between the DS scores. As a result, the distributions overlap more under the mismatch than the match condition and speaker discrimination is lower under the mismatch than

under the match condition. Left-shift among SS but not DS scores is a pattern that has been found elsewhere in the literature and a possible explanation has been given above. In casework, calibration would normally be applied, that means the forensically most relevant patterns are the ones illustrated in Figs. 3 and 4. These indicate that the system is more powerful under the match condition than under the mismatch condition, i.e. EER and Cllr are lower and there is a wider range of possible LLRs in the match condition. Casework can be conducted under both conditions if validation tests are available. Although speaker discrimination abilities are lower under mismatch, the system is still forensically useful, i.e. mismatch is not an exclusion criterion.

The results for the mismatch condition improved when Zoo plots were applied and the one speaker classified as "phantom" was removed from analysis. This improvement seems to indicate that screening for outlier speakers using Zoo plots can improve speaker discrimination and calibration.

The comparison of the unfamiliar condition [voice message] with the established condition [telephone], both in match condition, shows that both have similarly high speaker discrimination power (voice message: EER = 2.2%, telephone: EER = 2.3%). This means that if validation is available (as shown here), automatic speaker recognition can be performed equally well in comparisons based on [voice message] data as in comparisons based on [telephone] interception. It is not the case that [voice message] is a condition that is particularly challenging in itself. That would distinguish it from challenging conditions such as [Lombard] speech or [whisper] addressed in Kelly, Hansen (2021), where match condition performance within this [challenging] condition is lower than within a [neutral] condition, although mismatch between [neutral] and [challenging] conditions would be even poorer.

What if there was no validation data – could [voice message] cases be carried out with validation data based on [telephone] speech? Based on what we know from the presented tests, there is a difference in the score distributions of the conditions [voice message] and [telephone]. The difference is clear when looking at the raw scores but is much smaller when applying normalisation with case-relevant data to both, [voice message] and [telephone] data, respectively. If, for example, [voice message] comparisons are calibrated with [telephone] data, there will be a shift in LLRs that basically corresponds to the difference in the scores shown in Fig. 5 (when calibration is based on the raw scores [-norm, +CV]) or the scores shown in Fig. 6 (when calibration is based on the normalised scores [+norm, +CV]). A correct interpretation of the speaker comparison result would be hindered by the unknown extent of the shift. The hindrance would be low if normalisation was possible, i.e. if [voice message] data could be provided, but higher if not. It is a matter of judgement to define what constitutes a sufficiently small or too large interpretability bias. What should be done is to perform more tests under further unfamiliar conditions to get an impression about what constitutes large or small differences in the score distributions and whether results for some unfamiliar conditions could be interpolated based on previous tests. It is possible, for example,

that audio data from outdoor video recordings lead to larger score shifts than the ones observed here.

More data must be collected. Data collection should be done continuously by adding relevant casework data to the already existing corpus. This would allow for testing of additional languages (e.g. Turkish, Russian), recording conditions (e.g. video, interior surveillance) as well as mismatch scenarios for FASR application. Initial validation tests for language mismatch [Arabic – German] under [telephone] condition have already been carried out. However, more data is needed for valid and reliable interpretations. In addition, further languages and recording conditions in casework could become more frequent as new criminal networks emerge and technical possibilities develop.

Acknowledgements

The authors would like to thank Justin Hofenbitzer and Kjartan Beier for their contribution to the segmentation and anonymisation work as well as Linda Gerlach for discussing results gained from pre-tests and Finnian Kelly for providing important leads on Zoo plot screening and the interpretation of the results. We are grateful for the comments and suggestions gained from our colleague Almut Braun and from three anonymous reviewers to a previous version of this paper. In addition, we value each real forensic audio and video recording provided by colleagues from other departments within the BKA. Organisational support was also provided by the German-Israeli Cooperation on Counter-Terrorism Technology. This work was financially supported by the EU Internal Security Fund [grant number IZ25-5793-2018-34].

Bibliography

Bio-metrics 1.8 (2019). Performance Metrics Software, Oxford Wave Research Ltd., https://oxfordwaveresearch.com/products/bio-metrics/.

Boersma, P., Weenink, D. (2016). Praat: doing phonetics by computer. [Computer program] Version 6.0.17, retrieved 21 April 2016 from http://www.praat.org/.

Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J. & Niemi, T. (2015). *Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition*. Frankfurt: Verlag für Polizeiwissenschaft. https://enfsi.eu/wp-content/uploads/2016/09/guidelines_fasr_and_fsasr_0.pdf/ Accessed 05.12.2021.

Dunstone, T., Yager, T. (2009). *Biometric system and data analysis*. New York: Springer.

Fröhlich, A. (2017). Evaluation des forensischen Sprechererkennungssystems iVocalise anhand eines Zwillingsdatensets und eines Datensets, welches forensische Echtfallszenarien abbildet. Master thesis, Universität Zürich.

Gold, E., French, P. (2019). International practices in forensic speaker comparisons: second survey. In *International Journal of Speech Language and the Law*, 26, 1-20. http://dx.doi.org/10.1558/ijsll.38028

Hansen, J.H.L., Boril, H. (2018). On the issues of intra-speaker variability and realism in speech, speaker, and language recognition tasks. In *Speech Communication*, 101, 94-108. https://doi.org/10.1016/j.specom.2018.05.004

Jessen, M. (2018). Forensic voice comparison. In Visconti, J. (Ed. in collab. with Rathert, M.). *Handbook of communication in the legal sphere*. Berlin: Mouton de Gruyter, 219-255.

Kelly, F., Hansen, J.H.L. (2016). Score-aging calibration for speaker verification. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(12), 2414-2424. https://doi.org/10.1109/TASLP.2016.2602542

Kelly, F., Forth, O., Kent, S., Gerlach, L. & Alexander, A. (2019a). Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors. *Proceedings of the Audio Engineering Society (AES) Conference Audio Forensics*, Porto, Portugal, 18-20 June 2019, 1-7. https://www.aes.org/e-lib/online/search.cfm/ Accessed 05.12.2021.

Kelly, F., Fröhlich, A., Dellwo, V., Forth, O., Kent, S. & Alexander, A. (2019b). Evaluation of VOCALISE under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01). In *Speech Communication*, 112, 30-36. http://dx.doi.org/10.1016/j.specom.2019.06.005

Kelly, F., Hansen, J. (2021). Analysis and calibration of lombard effect and whisper for speaker recognition. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 927-942. https://doi.org/10.1109/TASLP.2021.3053388

Künzel, H.J. (2013). Automatic speaker recognition with crosslanguage speech material. In *The International Journal of Speech, Language and the Law*, 20(1), 21-44. http://dx.doi.org/10.1558/ijsll.v20i1.21

Morrison, G.S. (2011). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model – universal background model (GMM-UBM). In *Speech Communication*, 53, 242-256. https://doi.org/10.1016/j.specom.2010.09.005

Morrison, G.S. (2013). Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. In *Australian Journal of Forensic Sciences*, 45, 173-197. https://doi.org/10.1080/00450618.2012.733025

Morrison, G.S., Enzinger, E. (2016). Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) –introduction. In *Speech Communication*, 85, 119-126. http://dx.doi.org/10.1016/j.specom.2016.07.006

Morrison, G.S., Kelly, F. (2019). A statistical procedure to adjust for time-interval mismatch in forensic voice comparison. In *Speech Communication*, 112, 15-21. https://doi.org/10.1016/j.specom.2019.07.001

Morrison, G.S., Enzinger, E., Hughes, V., Jessen, M., Meuwly, D., Neumann, C., Planting, S., Thompson, W.C., Van der Vloed, D., Ypma, R.J.F., Zhang, C.,

Anonymous, A. & Anonymous, B. (2021). Consensus on validation of forensic voice comparison. In *Science and Justice*, 61, 229-309. https://doi.org/10.1016/j.scijus.2021.02.002

Nash, J. (2019). The effect of acoustic variability on automatic speaker recognition systems. PhD dissertation, University of York.

Rose, P. (2002). *Forensic speaker identification*. London: Taylor & Francis.

Shum, S., Dehak, N., Dehak, R. & Glass, J.R. (2010). Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification. *Proceedings of Odyssey*, Brno, Czech Republic, 28 June-1 July 2010, 76-82.

Solewicz, Y.A., Jessen, M. & van der Vloed, D. (2017). Null-hypothesis LLR: A proposal for forensic automatic speaker recognition. *Proceedings of Interspeech*, Stockholm, Sweden, 20-24 August 2017, 2849-2853.

Van Leeuwen, D.A., Brümmer, N. (2007). An introduction to application-independent evaluation of speaker recognition systems. In Müller, C. (Ed.). *Speaker classification I: Fundamentals, features, and methods*. Berlin: Springer, 330-353.

Van der Vloed, D., Kelly, F. & Alexander, A. (2020). Exploring the effects of device variability on forensic speaker comparison using VOCALISE and NFI-FRIDA: A forensically realistic database. *Proceedings of Odyssey*, Tokyo, Japan, 1-5 November 2020, 402-407. http://dx.doi.org/10.21437/Odyssey.2020-57

Vocalise 2.7 (2019). Automatic Speaker Recognition Software, Oxford Wave Research Ltd., https://oxfordwaveresearch.com/products/vocalise.