Automatic speaker recognition with variation across vocal conditions: a controlled experiment with implications for forensics

Vincent Hughes[1], *Jessica Wormald*[1], *Paul Foulkes*[1], *Philip Harrison*[1], *Finnian Kelly*[2], *David van der Vloe*d[3], *Poppy Welch*[1], *Chenzi Xu*[1]

[1]Department of Language and Linguistic Science, University of York, UK
[2]Oxford Wave Research, Oxford, UK
[3]Netherlands Forensic Institute, The Hague, The Netherlands

{vincent.hughes|jessica.wormald|paul.foulkes|philip.harrison|poppy.welch|chenzi.xu}@york.ac.uk, finnian@oxfordwaveresearch.com, d.van.der.vloed@nfi.nl

Abstract

Automatic Speaker Recognition (ASR) involves a complex range of processes to extract, model, and compare speaker-specific information from a pair of voice samples. Using heavily controlled recordings, this paper explores the impact of specific vocal conditions (i.e. vocal setting, disguise, accent guises) on ASR performance. When vocal conditions are matched, ASR performance is generally excellent (*whisper* is an exception). When conditions are mismatched, as in most forensic cases, we see an increase in discrimination and calibration error in some cases. The most problematic mismatches are those involving *whisper* and supralaryngeal vocal settings; these produce the greatest phonetic changes to speech. Mismatches involving *high pitch* also produce poor performance, although this appears to be driven by speaker-specific differences in articulatory implementation. We discuss the implications of the findings for the use of ASR in forensic casework and the interpretability of system output.

Index Terms: forensic speech science, automatic speaker recognition, voice quality

1. Introduction

Although much research into Automatic Speaker Recognition (ASR) has focussed on dealing with technical factors which could impact system performance (e.g. channel mismatch), recent work has demonstrated good performance in forensically realistic conditions when using case-specific data to optimise systems [1], [2].

Despite this, we still understand relatively little about why certain voices perform well or badly within automatic systems. This is particularly relevant in terms of system interpretability; following [3], we use the term 'interpretability' to refer to an understanding of a system's output. In forensics, this is especially important because a practitioner needs to understand whether the output of an ASR system is reasonable given the input and then explain this output to an end-user (e.g. a court). Further, because the conditions of each case are unique, we need a good understanding of what factors are important in order to collect and test relevant data when validating a system for a given case. Forensic cases often involve mismatches between recordings in terms of speaking style, context and technical characteristics, which can lead to substantial differences in the voices present in two samples, even where we know the voices are from the same speaker.

A small number of previous studies have attempted to assess what information ASR systems capture by fusing them with linguistic features [4], [5] or through the relatively small-scale analysis of the errors that a system produces [6], [7]. However, such work focuses on the output of the system with little control over the input speech. Our approach is to change the input in highly systematic ways in order to provide clearer interpretations of system output. Previous work [8] has explored same-speaker variability in x-vectors and scores generated by a state-of-the-art ASR system as a function of controlled variation in vocal conditions. The authors demonstrated that, when using *modal* voice as a baseline, the relationship between vocal conditions in the x-vector and score distributions reflects the degree of phonetic variability from *modal*: the conditions which resulted in the most global articulatory changes had the greatest difference in x-vectors and scores. Supralaryngeal changes in vocal setting (e.g. *backed tongue body*, *lowered larynx*), as well as *whisper,* resulted in the most divergent score and x-vector distributions. However, targeted, intermittent shifts of a range of features (as with accent guises) did not result in markedly different score and x-vector distributions.

The present paper expands on this work by exploring the impact of vocal variation with both same and different-speaker pairs, and considers the resultant LLRs and overall system performance after score calibration. In doing so, we seek to further our understanding of variability in ASR performance with vocal conditions which reflect the extremes of normal speech and those which are common in forensic casework. In turn, we hope to enable practitioners and courts to make the best use of ASR systems in casework.

2. Test data and scores

2.1 Test data

We report on a subset of material from a heavily controlled corpus, collected specifically to investigate the impacts of speaker and technical factors on ASR performance. The corpus includes variation in speaker, vocal condition, session and technical condition. Here, we are considering variation in vocal condition and speaker, whilst controlling for technical condition.

2.1.1 Participants

We report on data from six male phoneticians. Phoneticians were used as they are more effective at controlled vocal variation than lay speakers and are, in principle, able to independently vary certain settings. This meant that there was reduced between-speaker variability when varying vocal

10.21437/Interspeech.2023-443

conditions. The phoneticians had different levels of experience from PhD student to Emeritus Professor.

2.1.2 Vocal conditions

Each participant read the first two paragraphs of *The Rainbow Passage* in seventeen vocal conditions. The different conditions were selected to reflect large and small changes in segmental and suprasegmental vocal parameters. In part, conditions were also chosen to be representative of vocal variation found in forensic casework as a result of situational and stylistic factors as well as conscious disguise. Table 1 provides an overview of the vocal conditions included in the present study. For the accent guises and some of the miscellaneous conditions, participants were expected to vary multiple aspects of their vocal output (e.g. at both the segmental and suprasegmental levels). For the other conditions, our participants were asked to isolate and vary a single dimension holding others as fixed as possible.

Table 1: *Vocal conditions completed*

Baseline	MOD - Modal voice
Accent Guises	RPR - Received Pronunciation ACC - Non-standard guises, including: *Geordie, Manchester, NYC, Yorkshire*
Laryngeal	BRT - Breathy CRK - Creaky WHS - Whisper
Supralaryngeal	FTB - Fronted Tongue Body BTB - Backed Tongue Body RET - Retroflex LLX - Lowered Larynx
Miscellaneous	HIG - High pitch LOW - Low pitch FAS - Fast LIV - Lively MON - Monotone PEN - Pen between the teeth PIN - Pinched nose

2.1.3 Sessions

Each participant took part in three recording sessions which were at least a week apart. Within each session, each participant repeated each vocal condition three times. In this paper, we report on cross-session comparisons only (e.g. we have not considered within-session variability).

2.1.4 Technical conditions

All sessions were recorded in an anechoic chamber. Participants were seated at one end of the chamber throughout. Repetitions were simultaneously recorded in four technical conditions: headband microphone (DPA 4066 omnidirectional headset), near microphone (1m from participant), far microphone (2m from participant), and landline-to-VOIP call. Recordings were made in PCM WAV format with a 48kHz sample rate at 24 bits. For the purposes of the present study, only the headband microphone recordings were analysed. Individual repetitions of each vocal condition were extracted from within each session and are referred to throughout as a sample (i.e. 1 sample = 1 repetition of 1 condition in 1 session).

2.2. Comparisons and computation of scores

Comparisons were carried using the VOCALISE 2021 (version 3.0.0.1746) ASR system [9]. For each sample from each participant, we generated x-vectors [10] using the default x-vector model. These were generated from MFCCs, which were extracted on a frame-by-frame basis across the sample and then passed through a deep neural network (DNN). We then carried out same- (SS) and different-speaker (DS) comparisons for each sample in VOCALISE to generate scores using PLDA (e.g. x-vectors from each sample from each participant were compared to x-vectors from each other sample). For each matched and mismatched condition, we generated 618 SS and 3090 DS scores.

3. Calibration

3.1 Calibration data

Scores were calibrated using a subset of speakers from the DyViS corpus [11]. We included 20 speakers (ages: 18-25) who took part in both Task 3 and Task 5, which involved the speakers reading the same newspaper article 10-14 weeks apart. The recordings were made in a sound-treated recording studio. Although the content is different to the test data, the style and technical characteristics of the calibration set are comparable. There are some accent differences between the test and calibration data: one of the phoneticians is Scottish, and although none have particularly strong regional accents, the remaining five phoneticians do not speak Standard Southern British English as do those included in the DyViS corpus. Additionally, the age range represented in the test set was larger than that in the calibration set. However, the DyViS data were considered sufficiently well-matched for our purposes (as evidenced by very well calibrated log likelihood ratios in the *modal-modal* comparisons in section 4). Using a single calibration set also reflects a default scenario in casework where vocal properties of speakers in the calibration set are not controlled.

3.2 Score-to-LR conversion

Following the process described in section 2.2, 20 SS and 380 DS scores were generated for the calibration set using VOCALISE. Calibration was performed using the Bayesian model described in [12], [13], which reduces the magnitude of calibrated likelihood ratios (LRs) towards 1 (i.e. no support for prosecution or defence) when there is greater uncertainty. In our case, we used this model because the number of scores available to generate the calibration coefficients for score-to-LR conversion was relatively small [14]. In line with [13], we used Jeffreys uninformative priors and the scores from the calibration set to train the Bayesian model. The calibration coefficients from the model were then applied to the test scores to produce calibrated log LRs (LLRs).

3.3 Evaluation of performance

Tests were conducted using sets of matched and mismatched samples. System performance in each test was evaluated on the basis of calibrated LLRs, using the log LR cost function (C_{llr}) and its two constituent parts, the $C_{\mathrm{llr}}^{\mathrm{min}}$ and $C_{\mathrm{llr}}^{\mathrm{cal}}$. $C_{\mathrm{llr}}^{\mathrm{min}}$ is a measure of discrimination error which represents the lowest possible C_{llr} for each condition if the system were perfectly calibrated. $C_{\mathrm{llr}}^{\mathrm{cal}}$ is a measure of calibration error (although see [15]) and reflects how well suited the calibration set is for the test data. C_{llr} is the sum of the $C_{\text{llr}}^{\text{min}}$ and the $C_{\text{llr}}^{\text{cal}}$ and a value of above 1 means the system is not providing meaningful information for separating SS and DS pairs (this may be due to either discrimination or calibration error, or a combination). The use of high-quality, channel-matched samples is expected to produce very good system performance; considerably better than what would be expected in forensic casework. However, our interest is in relative performance across the matched and mismatched conditions. Further, by using samples of optimal quality, we remove confounding variables allowing us to better isolate the effects of vocal condition on performance.

4. Results

4.1 Overall performance

Figure 1 displays the $C_{\text{llr}}^{\text{min}}$ (x-axis) and $C_{\text{llr}}^{\text{cal}}$ (y-axis) for all condition pairs. For the mismatched conditions (each condition to each other condition), each label on the plot represents the median value for that condition compared with all other conditions. For example, WHS on the mismatched bottom plot reflects the median $C_{\text{llr}}^{\text{min}}$ and $C_{\text{llr}}^{\text{cal}}$ values for all *whisper-other* test sets for all speakers across all sessions.

In general, the system is performing very well for matched conditions, with all vocal conditions generally clustering around the bottom left corner with a low $C_{\text{llr}}^{\text{min}}$ (around 0) and $C_{\text{llr}}^{\text{cal}}$ (< 0.25). Thus, when technical conditions are held constant, and the vocal condition is matched across recordings, the system is able to distinguish SS and DS pairs very well. Although this is what we might expect (especially for *modal*), it is reassuring that we see no marked effects for any of the other conditions. The exception to this is the *whisper* condition, which has slightly poorer discrimination error ($C_{\text{llr}}^{\text{min}}$ = 0.07) and considerably poorer calibration error ($C_{\text{llr}}^{\text{cal}}$ = 0.91). Unsurprisingly, the mismatched conditions generate poorer performance than the matched conditions. Figure 1 also highlights that the mismatched conditions produce considerably more variability in system performance. Mismatched conditions involving *modal*, *monotone*, *low pitch*, *fast*, *Received Pronunciation* and all of the other accent guises have relatively little effect on performance, with low values for both the $C_{\text{llr}}^{\text{min}}$ and $C_{\text{llr}}^{\text{cal}}$. However, while the $C_{\text{llr}}^{\text{min}}$ is below 0.6 for all condition pairs, the $C_{\text{llr}}^{\text{cal}}$ is above 1 for many mismatched conditions, particularly, *whisper*, *high pitch*, *pinched nose*, *pen between the teeth*, and all but one of the Supralaryngeal conditions (*retroflex*, *backed tongue body*, and *lowered larynx*). For these mismatched conditions, the poor calibration is driven by a general left-ward shift in LLRs. This means that SS LLRs shift towards contrary-to-fact support for the different-speaker proposition, while DS LLRs shift towards even stronger support for the different-speaker proposition. Given that the calibration data used is modal only, calibration error may be reduced with matched condition calibration data.

4.2 Which vocal conditions matter?

4.2.1 Modal to other

In this section we consider pairs of vocal conditions where one condition is *modal*. This allows us to assess the impact of each vocal condition relative to a baseline. Only three *modal* mismatch pairs have a C_{llr} greater than 1: *lowered larynx*, *high pitch*, and *whisper*. *Modal-high pitch*, and *modal-whisper* are

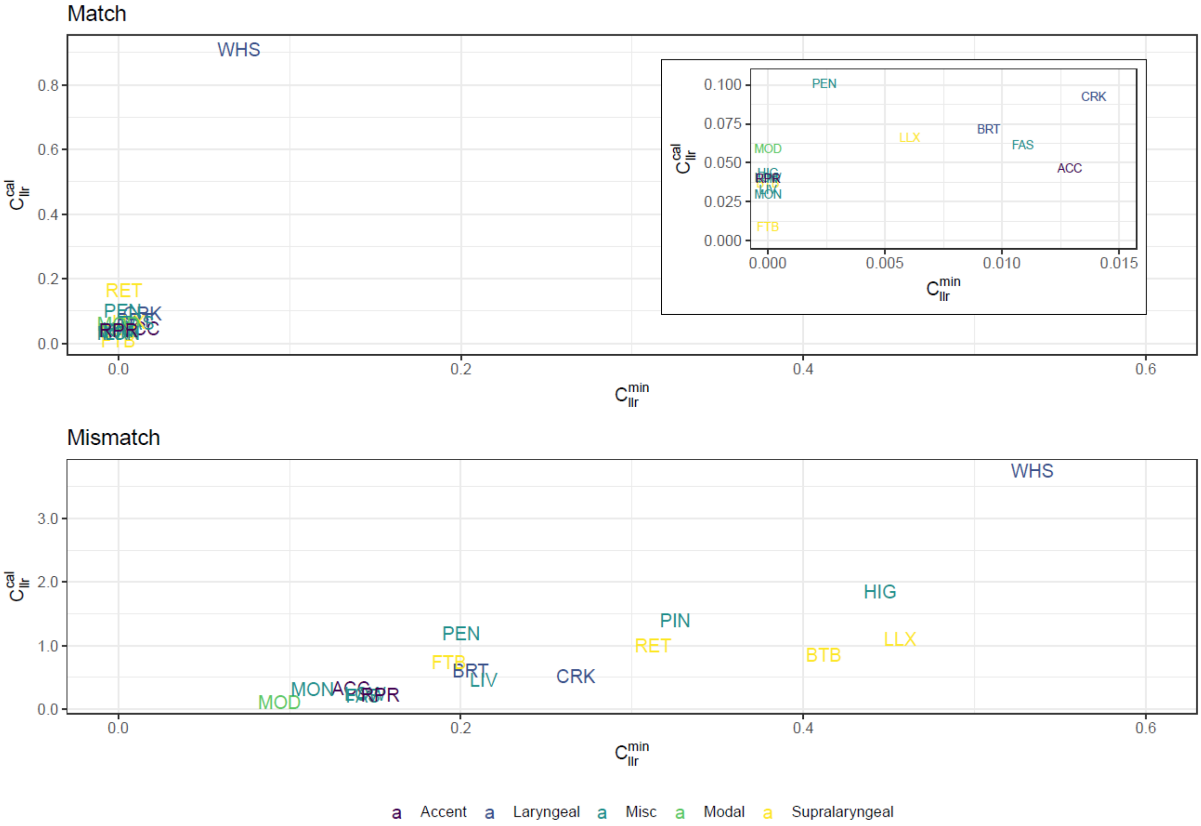

Figure 1: C_{llr}^{min} *(x-axis) and* C_{llr}^{cal} *(y-axis; note different scales) for each condition pair. Matched condition pairs are in the top plot, mismatched condition pairs are in the bottom plot and are based on the median across all tests.*

mismatches which could realistically occur in a forensic case. The effect of *whisper* is partly a calibration issue, since the calibration data does not include any whispered speech. We discuss *whisper* in more detail in section 4.2.2, and *high pitch* in section 4.3.

Four *modal-other* condition pairs have a C_{llr} between 0.5 and 1: *pen between the teeth*, *pinched nose*, *backed tongue body* and *retroflex*. The first two of these could be employed as potential disguise strategies in forensic cases and thus we might encounter mismatches of this kind. The remaining condition pairs have C_{llr} values below 0.5. These condition pairs involve variation from *modal* in terms of speech rate, lowering of pitch, targeted segmental variation (e.g. *accent guises*), or variation in phonation. For these condition pairs, all of which are also forensically realistic, the system is able to cope with the increase in within- and between-speaker variability. These vocal condition pairs aren't having a marked effect on system performance.

4.2.2 Whisper to other

The mismatch condition pairs which include *whisper* have the poorest overall performance, both in terms of discrimination and calibration error. Discrimination performance is good in *whisper*-*whisper* comparisons, therefore calibration loss is a major contributor to the error. In mismatched conditions, all condition pairs have a C_{llr} above 2. *Whisper* is the only condition to have a complete absence of voicing: whilst there is turbulent airflow at the glottis, there is no periodic vibration or regular closure of the folds. This lack of voicing has substantial effects on the speech signal that the ASR system is clearly sensitive to.

4.3 Between-speaker variability

In addition to variability in overall system performance within and between conditions, we also found effects driven by between-speaker variability. In many cases, this was phonetically explainable in terms of the way in which, or degree to which, the speakers produced some of the vocal conditions. Here we focus on one condition which exemplifies this issue; *high pitch*.

As illustrated in Figure 1, with the exception of *whisper*, mismatch condition pairs involving *high pitch* have the highest $C_{\mathrm{llr}}^{\mathrm{min}}$ and $C_{\mathrm{llr}}^{\mathrm{cal}}$. However, there was a large amount of between-speaker variability in terms of how speakers produced 'high pitched' speech. To assess the effects of this on system performance, we also calculated C_{llr} by-speaker. The speaker (P1) with the highest median f_0 in the *high pitch* condition (and who displayed the largest f_0 median difference between the *modal* and *high pitch* conditions) produced the highest C_{llr} of the six speakers. However, both auditory analysis of the recordings and comments from the speaker himself revealed that *high pitch* was achieved not only through an increase in vocal fold vibration and laryngeal tension, but also by raising the larynx. With the exception of this speaker, no clear relationship between change in f_0 and C_{llr} was found, although the sample size is extremely small. Indeed, other speakers (e.g. P4 and P6) reported actively attempting to raise their pitch only through an increase in vocal fold vibration rather than any other compensatory articulatory changes. This suggests that high pitch related to increased vocal fold vibration, e.g. in Lombard speech, itself may not have substantial effects on ASR performance. The issue for ASR is when extreme high pitch is achieved through a combination of vocal effects: increased f_0, raising of the larynx and shortening of the supralaryngeal vocal tract, sometimes leading to falsetto. This combination of vocal effects has a substantial impact on the spectrum, similar in magnitude to those seen in the *lowered larynx* and *backed tongue body* conditions in Figure 1.

5. Discussion

Overall, when vocal conditions are matched we see there is little impact on ASR performance. Mismatched vocal conditions result in more variable and generally poorer performance.

When considering the interpretability of ASR systems, we have demonstrated that the variability observed in the mismatch conditions is, to some extent, phonetically predictable. Principally, supralaryngeal conditions are generally most problematic because of the large-scale and long-term changes they produce. Accent guises, and changes to speech rate, pitch, and phonation generally have little effect. This also highlights that varying multiple aspects of vocal output at both segmental and suprasegmental levels (e.g. an accent guise) can have a less substantial impact than changes in a single vocal condition which results in wider articulatory shifts and overall spectral changes (e.g. *whisper*, or *lowered larynx*).

Whisper can be considered a special case because of the complete absence of voicing (see also [16]). The findings here highlight that forensic cases which involve a sample of whispered speech should endeavour to use calibration data which also includes whispered speech. Whilst specifically applicable to *Whisper*, it is also likely that more tailored calibration data would lead to considerable improvements in all mismatched conditions.

Although we have highlighted condition pairs which can be problematic, this paper also demonstrates that the ASR system can deal with a number of mismatched conditions very well when all other factors are held constant. When exploring variation from *modal*, we observed similar patterns to previous work in [8] in that modifications to speaking rate, targeted segmental variation (e.g. *accent guises*), most phonation deviations, and lowering of pitch (in the absence of raising/lowering the larynx) do not have a marked impact on system performance. Additionally, even for features such as *pen between the teeth*, or *pinched nose*, the C_{llr} is still below 1.

6. Conclusions

In this paper we have demonstrated that when vocal conditions are matched, ASR performance is generally excellent. In addition, mismatch conditions which do not result in marked spectral changes are also not problematic for the system. Some mismatch conditions can impact on system performance - both in terms of discrimination and calibration error. However, these impacts are generally interpretable when taking into account the degree of spectral change related to those vocal changes. Overall, there is still more to do to understand how such issues manifest when conditions are more challenging and more reflective of unique and variable forensic cases.

7. Acknowledgements

This work was funded by ESRC grant ES/W001241/1.

8. References

[1] G. S. Morrison and E. Enzinger, “Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Introduction,” *Speech Commun.*, vol. 85, pp. 119–126, Dec. 2016.

[2] G. Morrison and E. Enzinger, “Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01)–Conclusion,” *Speech Communication*, 2019,

[3] Y. Pruksachatkun, M. Mcateer, and S. Majumdar, *Practicing Trustworthy Machine Learning*. O’Reilly Media, Inc., 2023.

[4] E. Enzinger, C. Zhang, and G. S. Morrison, “Voice source features for forensic voice comparison – an evaluation of the Glottex software package,” in *Proceedings of Odyssey 2012: The Language and Speaker Recognition Workshop*, Singapore, 2012, pp. 78–85.

[5] V. Hughes, P. Foulkes, and S. Wood, “Formant dynamics and durations of um improve the performance of automatic speaker recognition systems,” in *Proceedings of the 16th Australasian Conference on Speech Science and Technology (ASSTA)*, Australia, 2016, pp. 249–252.

[6] J. González-Rodriguez, J. Gil, R. Perez, and J. Franco-Pedroso, “What are we missing with i-vectors? A perceptual analysis of ivector-based falsely accepted trials,” in *Odyssey 2014: The Language and Speaker Recognition Workshop*, 2014, pp. 33–40.

[7] V. Hughes, P. Harrison, P. Foulkes, J. P. French, C. Kavanagh, and E. San Segundo, “Mapping across feature spaces in forensic voice comparison: the contribution of auditory-based voice quality to (semi-) automatic system testing,” in *Proceedings of Interspeech*, Stockholm, Sweden, 2017, pp. 3892–3896.

[8] J. Wormald, P. Foulkes, P. Harrison, V. Hughes, F. Kelly, D. van der Vloed, P. Welch and C. Xu, “Sensitivity of x-vectors and automatic speaker recognition scores to vocal variation,” in *Proceedings of ICPhS*, Prague, 2023.

[9] F. Kelly, O. Forth, S. Kent, L. Gerlach, and A. Alexander, “Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors,” in *Audio Engineering Society Conference: 2019 AES International Conference on Audio Forensics*, 2019.

[10] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-Vectors: Robust DNN Embeddings for Speaker Recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Apr. 2018.

[11] F. Nolan, K. McDougall, G. de Jong, and T. Hudson, “The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research,” *International Journal of Speech, Language and the Law*, vol. 16, no. 1, pp. 31–57, 2009.

[12] N. Brümmer and A. Swart, “Bayesian calibration for forensic evidence reporting,” in *Interspeech 2014*, Sep. 2014. doi: 10.21437/interspeech.2014-90.

[13] G. S. Morrison and N. Poh, “Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios/Bayes factors,” *Sci. Justice*, vol. 58, no. 3, pp. 200–218, May 2018.

[14] B. X. Wang and V. Hughes, “Reducing uncertainty at the score-to-LR stage in likelihood ratio-based forensic voice comparison using automatic speaker recognition systems,” in *Proceedings of Interspeech*, Incheon, Korea, 2022, pp. 5243–5247.

[15] G. S. Morrison, “In the context of forensic casework, are there meaningful metrics of the degree of calibration?,” *Forensic Sci Int Synerg*, vol. 3, p. 100157, Jun. 2021.

[16] F. Kelly and J. H. L. Hansen, “Analysis and Calibration of Lombard Effect and Whisper for Speaker Recognition,” *IEEE/ACM Trans Audio Speech Lang Process*, vol. 29, pp. 927–942, Jan. 2021.