HHS Public Access

Author manuscript

IEEE/ACM Trans Audio Speech Lang Process. Author manuscript; available in PMC 2022 June 30.

Published in final edited form as:

IEEE/ACM Trans Audio Speech Lang Process. 2021 ; 29: 927–942. doi:10.1109/taslp.2021.3053388.

Analysis and Calibration of Lombard Effect and Whisper for Speaker Recognition

Finnian Kelly,

John H.L. Hansen [Fellow, IEEE]

Center for Robust Speech Systems (CRSS), University of Texas at Dallas, Richardson, TX 75083-0688 USA

Abstract

Variations in vocal effort can create challenges for speaker recognition systems that are optimized for use with neutral speech. The Lombard effect and whisper are two commonly-occurring forms of vocal effort variation that result in non-neutral speech, the first due to noise exposure and the second due to intentional adjustment on the part of the speaker. In this article, a comparative evaluation of speaker recognition performance in non-neutral conditions is presented using multiple Lombard effect and whisper corpora. The detrimental impact of these vocal effort variations on discrimination and calibration performance on global, per-corpus, and per-speaker levels is explored using conventional error metrics, along with visual representations of the model and score spaces. A non-neutral speech detector is subsequently introduced and used to inform score calibration in several ways. Two calibration approaches are proposed and shown to reduce error to the same level as an optimal calibration approach that relies on ground-truth vocal effort information. This article contributes a generalizable methodology towards detecting vocal effort variation and using this knowledge to inform and advance speaker recognition system behavior.

Keywords

Calibration; Lombard effect; speaker recognition; speech processing; whisper/vocal effort

I. Introduction

A MAJOR challenge for speaker recognition systems operating in real-world conditions is to effectively overcome variability in the speech signal. There are many factors affecting speech presented to a speaker recognition system; these can be loosely grouped into factors independent of the speaker (extrinsic factors), and factors dependent on the speaker (intrinsic factors). Extrinsic factors include environmental noise, room acoustics, and the effects of the microphone and transmission channel. Intrinsic factors include the speaking style and conversational context, the emotional state of the speaker, changes to the speaker's health, and the longer-term impact of aging.

(Corresponding author: John H. L. Hansen. john.hansen@utdallas.edu).

Developments in speaker recognition technology continue to be driven by performance improvements in the presence of extrinsic variability (e.g., the focus on challenging audio From video conditions in NIST Speaker Recognition Evaluations 2018–19 [1]); this has resulted in speaker recognition technology reaching a level of maturity for applications involving general extrinsic variability. However, in the application of speaker recognition to unconstrained domains such as forensics (where, for example, speech may be under stress, highly emotional, or deliberately altered), there remains a real need to further understand and address the impact of intrinsic variability.

The study of intrinsic variability presents challenges in terms of data collection; for example, real emotional speech is difficult (or perhaps impossible) to collect in a way that is both controlled and ethical, and aging speech requires a long period of time to elapse if it is to be obtained in a controlled way. As a result, research addressing intrinsic variability typically leverages smaller, curated datasets, making generalization difficult. In this study, we consider two specific modes of speech resulting from vocal effort variation, namely whispered speech, and speech produced under the Lombard effect (in the presence of noise). These modes affect the speech signal in very different ways, but are both commonly-occurring and relevant to real-world applications. To improve generalization of results and their interpretation in this study, we present analyses across multiple corpora for each speech mode.

The Lombard effect (LE) [2], [3] (also referred to as the Lombard reflex), refers to the involuntary tendency of a speaker to alter their vocal effort in the presence of environmental noise so that intelligible communication is preserved. Speech produced under the Lombard effect differs from neutral speech in several ways, including increased intensity, pitch, glottal spectral tilt, and F1 [3]–[5], along with a lengthening of vowel segments and a shortening of silence segments [3], [5]. The type and level of noise inducing the Lombard effect has also been shown to affect the extent of the speech production changes [3]. The Lombard effect occurs frequently in everyday speech, in a noisy room, or over a poor telephone connection, for example. For this reason, it is important to consider its effect on speaker recognition. The Lombard effect has previously been shown to impact the performance of a range of speech technology applications [3], [6]. Efforts to compensate for the Lombard effect have focused primarily on speech recognition applications [4], [5], [7]–[9]. Previous speaker recognition studies have considered detection and integration of Lombard speech into system training for GMM (Gaussian Mixture Model - Universal Background Model) [3] and i-vector PLDA (Probabilistic Linear Discriminant Analysis) [10] systems as a means of compensating for performance loss. In our previous work [11], the Lombard effect was shown to have a detrimental impact on i-vector PLDA speaker recognition, in a way that was reflective of the noise type and level inducing the Lombard effect. It was demonstrated that the reference noise level could be integrated into score calibration to improve discrimination and calibration performance.

Whisper is a common mode of low vocal effort speech, generally produced with the aim of maintaining intelligibility for an intended listener, while restricting intelligibility for others. It is therefore commonly utilized in scenarios where a speaker wishes to conceal or disguise their identity [12], or communicate information discretely [13]. Whisper differs significantly

from neutral speech (modal speech produced with normal vocal effort). Whisper is produced without vibration of the vocal folds, and therefore the signal contains no periodic excitation (this distinguishes whisper from soft or low vocal effort speech). Additionally, the center frequencies and bandwidths of formants generally increase in whisper, while overall signal energy decreases [14], [15]. From a forensic perspective, the nature of whisper as a means to conceal identity or information give it particular relevance. Aside from forensic speaker recognition contexts, the ability to discreetly communicate with commercial applications of speech or speaker recognition is of growing interest (e.g., Amazon Alexa, an in-home speech-controlled virtual assistant, supports a 'whisper mode').Previous studies[15]–[18] have established that whisper-neutral speech comparisons significantly degrade performance of conventional speaker recognition frameworks relative to neutral-neutral comparisons, and have proposed various front-end feature modifications to address this issue. In our previous work [19], whisper was shown to have a detrimental impact on i-vector PLDA speaker recognition. A cross-corpus i-vector-based whisper detection scheme was introduced to select calibration parameters for each test comparison, leading to an improvement in overall discrimination and calibration performance.

Whisper and Lombard speech deviate from neutral speech in very different ways, but have both been shown to negatively impact speaker recognition performance. The goal of this study is therefore to evaluate and compare the effect of these speech modes with a consistent speaker recognition system and testing protocol, and to propose score calibration schemes that can be successfully applied to comparisons involving these speech modes. The process of score calibration involves transforming scores output by a speaker recognition system to have some desirable properties: calibrated scores can be used to make reliable decisions with a fixed decision threshold, or can be interpreted directly as log likelihood ratios [20], which are suitable for forensic applications. In order to achieve effective score calibration, the presence of (extrinsic or intrinsic) variability must be taken into account. For example, Fig. 1 shows distributions of the same-speaker and different-speaker scores for comparisons in two different conditions, where Condition A consists of typical operating conditions, and Condition B contains some additional variability. The system in Fig. 1(a) applies the same calibration transformation to the scores in both conditions, resulting in poor overall calibration (reflected in the misalignment of the two sets of score distributions), whereas the system in Fig. 1(b) applies an appropriate calibration transformation to each of the conditions independently, resulting in good overall calibration (reflected in the alignment of the two sets of score distributions).

In the context of this study, Condition A is representative of comparisons with neutral speech, and Condition B is representative of comparisons between neutral speech and whisper or Lombard effect speech. Our goal is to automatically detect when the system is operating under Condition B, and to adjust the calibration transformation appropriately. An advantage of this approach is that it can be applied in a way that is independent of the speaker recognition framework (i.e. the feature extraction, speaker modeling, and speaker comparison algorithms), and could potentially be generalized to other sources of variability beyond Lombard effect and whisper.

A recent study [21] evaluated the effect of three vocal effort levels (low, normal, high) on speaker recognition performance, and considered condition-specific PLDA models and trial-based calibration [22] to improve discrimination and calibration performance. The aim of our study is similar, but our approach is focused on the use of a non-neutral speech detector to inform calibration schemes that are system-independent and corpus-independent. We also note that several other recent studies [23], [24] have focused on the effect of extrinsic variability on calibration performance, and have considered several ways of dealing with the associated drop in performance.

The contributions of this study include a series of speaker recognition experiments with three Lombard effect speech corpora and three whispered speech corpora using an i-vector system [25] and a Deep Neural Network (DNN) based x-vector embeddings system [26] (there is variability in language, accent and speech content across the corpora; however, we constrain the experiments to fixed-duration comparisons of close microphone read speech). The application of a Support Vector Machine (SVM) to detect non-neutral speech from x-vectors, and the use of the detector output to inform score calibration, are shown to lead to improvements in calibration and discrimination performance in the presence of Lombard effect and whisper vocal effort. This study consolidates and extends our previous work [11], [19] by evaluating both Lombard effect and whisper within the same speaker recognition framework, and by proposing generalizable calibration schemes that are applicable to both speech modes, and across different corpora.

II. Speech Data

The availability of suitable speech data has been a limiting factor in the study of intrinsic variability in speaker recognition. In this study, several independently-collected corpora are subjected to the same analysis, enabling some general conclusions to be drawn. All corpora were designed and collected to study the specific vocal effort variations of Lombard effect and whisper. The following sections introduce each of the datasets in detail, and a summary is provided in Table I.

A. UT-SCOPE: Lombard Effect Speech

UT-SCOPE (Speech under COgnitive and Physical stress and Emotion) [27] is a corpus containing a range of intrinsic speaker variability, including Lombard effect and neutral speech recordings produced in controlled conditions (UT-SCOPE additionally contains recordings of speech produced under cognitive and physical stress, not considered in this study). The Lombard speech portion of the corpus used in this study (UT-SCOPE-LE) contains English recordings of 30 (24 female, 6 male) native US-English speakers. The Lombard effect was induced by playback of several noise types at different levels using open-air headphones. The noise types considered were: large crowd noise, highway noise (recorded inside a car traveling at 65 mph on a highway with the windows half-open), and pink noise. Each noise type was presented at three levels: crowd and highway at 70, 80 and 90 dB-SPL, and pink at 65, 75 and 85 dB-SPL.Neutral speech was recorded under the same conditions, with no noise presentation. Recordings were made in an acoustically-isolated booth with a Shure SM-10 A close-talk microphone, along with a desk-top microphone

and two distant microphones (only the close-talk mic. recordings are considered here). Lombard effect and neutral speech recordings were made within the same session as blocks of read and spontaneous speech (only the read portion is considered in this study). The read Lombard speech blocks consisted of 20 TIMIT sentences and five repetitions of 10 digits in each noise condition; the read neutral speech consisted of 100 TIMIT sentences and five repetitions of 10 digits with no noise presentation. As the Lombard Effect was induced by playback of noise though headphones, the noise signal is not present in the speech recordings. The headphones were open-air, and so there was no occlusion effect. The hearing of the speakers was screened in advance using a pure-tone audiometric test.

B. Pool-2010: Lombard Effect Speech

Pool-2010 [28] is a corpus of Lombard effect and neutral speech recordings produced in controlled conditions. The corpus contains German recordings of 100 native German speakers. A subset of 100 male speakers were selected, based on recommendations in [28]. The Lombard effect was induced by playback of white noise over headphones at 80 dB-SPL. Neutral speech was recorded under the same conditions, with no noise presentation. Recordings were made with a head-mounted close-talk microphone in an acoustically-isolated booth. Lombard effect and neutral speech recordings were made within the same session, as blocks of read and spontaneous speech (only the read portion is considered in this study). Read speech consisted of a read passage (the German version of the ‘North Wind and the Sun’) in each condition. In addition to the microphone recordings, telephone recordings were simultaneously acquired for the same speech tasks. Only the microphone recordings are considered in this study. As the Lombard Effect was induced by playback of noise though headphones, the noise signal is not present in the speech recordings. The hearing of the speakers was not screened; however, it is noted that none of the speakers had ‘noticeable voice or speech disorders’ [28].

C. Lombard-Grid: Lombard Effect Speech

Lombard-Grid [29] is an audio-visual corpus of Lombard effect and neutral speech recordings produced in controlled conditions.[1] The corpus contains English recordings of 54 (30 female, 24 male) native British-English speakers. The Lombard effect was induced by playback of speech-shaped noise over headphones at 80 dB-SPL. Neutral speech was recorded under the same conditions, with no noise presentation. Recordings were made in an acoustically-isolated booth with a C414 B-XLS AKG condenser microphone at a distance of 30 cm. Lombard effect and neutral speech recordings were made within the same session as alternating blocks of 10 read sentences. Frontal and side video was recorded alongside the audio (only the audio portion is considered in this study). As the Lombard Effect was induced by playback of noise though headphones, the noise signal is not present in the speech recordings. The hearing of the speakers was screened in advance using a pure-tone audiometric test.

[1]The corpus is available for download at: http://spandh.dcs.shef.ac.uk/avlombard/ (last accessed 17th November 2020).

D. UT-VocalEffort II: Whisper

UT-VocalEffort (VE) II [30] is a corpus of whispered and neutral speech recordings produced naturally in controlled conditions. A subset of the corpus consisting of English recordings of 62 (42 female, 20 male) native US-English speakers was used in this study. Recordings were made in an acoustically-isolated booth with a Shure Beta-53 headset microphone. Whisper and neutral speech recordings were made within the same session, in both read and spontaneous conversational contexts (only the read portion is considered in this study). The read speech content consists of 41 TIMIT sentences, produced in neutral and whisper modes in an alternating fashion. All subjects confirmed no history of hearing or speech impairment.

E. CHAINS: Whisper

CHAINS (CHAracterizing INdividual Speakers) [31] is a corpus of whispered and neutral speech recordings produced naturally in controlled conditions.[2] The corpus contains English recordings of 36 speakers (16 females, 20 males). The majority of the speakers (28) are Irish-accented, and the remainder of the speakers are US or U.K.-accented. Neutral speech recordings were collected with a Neumann U87 condenser microphoneinan acoustically-isolated booth. Whispered speech recordings were collected in a session about 2 months later, with an AKG C420 headset condenser microphone in a quiet office environment. Recordings of both read and spontaneous speech were made (only the read portion is considered in this study). The read speech content consists of 4 read passages and 33 read sentences (24 from TIMIT, 9 from CSLU), produced within each of the separate neutral and whisper recording sessions.

F. wTIMIT: Whisper

wTIMIT (whispered TIMIT) [32] is a corpus of whispered and neutral speech recordings produced naturally in controlled conditions. The corpus contains speech from 48 speakers (24 females, 24 males). The recordings are in English, with 20 Singapore-accented speakers and 28 US-accented speakers. Neutral speech recordings were collected with a MX-2001 directional microphone in an acoustically-isolated booth. Whisper and neutral speech recordings were made within the same session, in an alternating fashion. Speech consisted of 450 read sentences from the TIMIT corpus [33] produced in both whisper and neutral modes, and spoken in batches of 50.

G. Notes on Speech Data

Although these copora were collected independently, they share many consistencies. In the selection of data subsets and the design of experiments, we have imposed some additional constraints to ensure the corpora are more comparable:

- All recordings used for experiments were made with a close-talk or directional microphone in an acoustically-isolated booth, with the exception of CHAINS whisper recordings, which were made in a quiet office.

[2]The corpus is available for download at: http://chains.ucd.ie (last accessed 17th November 2020).

- All recordings used for experiments consisted of read sentences, with the exception of Pool2010, which consisted of a read passage. There is variation in the total speech duration across corpora; however, duration consistency is enforced in our speaker recognition experiments.
- The number of neutral and non-neutral recordings is approximately equal for all corpora, with the exception of UT-SCOPE-LE (due to the multiple Lombard conditions); this imbalance is accounted for at the analysis stage.
- The corpora are relatively gender balanced, with the exception of Pool2010 (all male) and UT-SCOPE-LE (80% female)
- All Lombard speech collections used noise at 80 dB-SPL as a stimulus (UT-SCOPE-LE additionally uses noise at lower and higher levels), and all collected Lombard speech recordings are clean (i.e. do not contain the noise stimulus).
- All recordings were down sampled to 8 kHz prior to analysis
- All recordings used for experiments were made within the same session (i.e. recording event), with the exception of CHAINS whisper recordings, which were made in a separate session two months later. This is considered in the interpretation of results, as comparison of speech samples recorded within the same session typically produce higher speaker recognition scores than would be observed with comparisons across different sessions.

III. Speaker Recognition Systems

This section details the speaker recognition systems used for experiments. Our primary system is an x-vector PLDA system based on DNN embeddings [26]. For comparison, and continuity with our previous studies [11], [19], an i-vector PLDA system [25] is additionally used for a subset of our experiments.

A. i-vector System

An i-vector PLDA system [25], [45] was trained on NIST SRE 2004–2010 collections. At the system front-end, 15 Mel-frequency cepstral coefficients (MFCCs) were extracted over 20 ms windows at 10 ms intervals, and were appended with first- and second-order derivatives (delta and delta-delta coefficients). Combo-SAD [46] speech activity detection (SAD) was applied to remove non-speech frames. Cepstral mean subtraction (CMS) [47] was applied globally (i.e. the mean of the entire sample was subtracted from each frame). A gender-independent UBM (Universal Background Model) of 1024 components and a 400-dimensional total variability matrix were trained with the full training set. i-vectors were post-processed by mean and length normalization, and whitening [48]. The i-vector dimensionality was reduced to 200 using Linear Discriminant Analysis (LDA), and the same-speaker and different-speaker i-vector distributions were modeled with PLDA using the full training set.

B. x-vector System

An x-vector PLDA system [26] with a Kaldi network architecture was trained using the NIST SRE 2016 recipe.[3] At the front-end, 23 MFCCs were extracted over 25 ms windows at 10 ms intervals, and energy-based SAD was applied to remove non-speech frames. CMS was applied with a sliding window of 3 seconds across the speech sample. The speaker embeddings extractor used is a feed-forward DNN: the first five DNN layers operate on speech frames with an increasing temporal context (i.e. an increasing time span between first and last input frame). The frame-level layers are followed by a statistics pooling layer that aggregates information across all frames of the input speech sample. This is followed by two hidden layers, before a final softmax output layer. The first hidden layer, of 512 dimensions, is taken as the speaker embedding, or x-vector. The x-vector dimensionality was reduced to 150 with LDA, and the same-speaker and different-speaker x-vector distributions were modeled with PLDA using the full training set.

IV. Speaker Recognition Under Vocal Effort Variation

A set of baseline speaker recognition experiments were conducted for each of the corpora to assess the impact of Lombard effect and whisper vocal effort variation on the performance of i-vector and x-vector systems. The aim of these initial experiments was to establish the performance for matched (neutralneutral, Lombard-Lombard, and whisper-whisper) and mismatched (Lombard-neutral and whisper-neutral) comparisons in a comparable way across the various corpora.

A. Discrimination Performance

As the corpora largely consist of short sentence utterances, the content of each was restructured to control for duration (there is a strong relationship between duration and speaker recognition performance [49], [50]). Recordings were concatenated into one non-neutral (i.e. Lombard speech or whisper) and one neutral recording of read speech per-speaker. The concatenated recordings were then segmented into fixed-duration chunks of 10 seconds net speech (i.e. disregarding non-speech and silence). Speech activity labels from Combo-SAD [46] were used to segment the recordings in this way. A minimum of one chunk and a maximum of five chunks were extracted from each recording; for the case where more than five chunks were available, they were chosen to be maximally distributed throughout the recording. From our previous investigation of the UT-VE II and CHAINS whisper corpora [19], it was observed that i-vector speaker recognition performance tended toward 0% Equal Error Rate (EER) at durations above 20 seconds. By limiting the duration of all test samples to 10 seconds, we therefore focus on a more challenging short-duration task.

For each corpus, all possible same-gender recording chunks were compared using both i-vector and x-vector systems. The discrimination performance was measured for matched and mismatched comparison conditions by calculating the EER on the relevant scores. Table II provides the EERs for all conditions within each Lombard effect and whisper corpus.

[3] https://david-ryan-snyder.github.io/2017/10/04/model_sre16_v2.html (last accessed 17th November 2020).

For the 'ALL' condition, a weighted EER [51] was calculated in order to balance the contribution of matched and mismatched comparisons. In calculating a weighted EER, the contribution of comparisons originating from a particular condition (i.e. matched or mismatched) was weighted by the inverse of the total number of comparisons for that condition. The weighting of error metrics in this way was proposed for evaluations with comparisons in multiple recording conditions, and has also been applied in the case of a variable number of comparisons per-speaker [52].

Referring to Table II, it is clear that mismatched comparisons of neutral vs. non-neutral speech(N-NN) degrade discrimination performance relative to neutral-neutral comparisons (N-N), and that whisper has a much greater negative impact than Lombard effect speech. The large performance improvement offered by the x-vector system across all conditions is also clear. It is important to note that while the improvement from i-vector to x-vector for matched (N-N, NN-NN) conditions is apparent, the significant loss in performance for mismatched (N-NN) conditions persists for both.

In absolute terms, the x-vector EERs for mismatched Lombard effect vs. neutral comparisons are low (a maximum EER of 3.62% for Lombard-Grid); however, these EERs still represent a large increase relative to the corresponding neutral-neutral comparisons. Absolute discrimination performance of the x-vector system is poor for mismatched whisper vs. neutral comparisons, with minimum EERs of ≈20%.

It is interesting that matched comparisons of non-neutral speech (i.e. Lombard effect vs. Lombard effect, and whisper vs. whisper (NN-NN)), result in lower EERs than mismatched comparisons of neutral and non-neutral speech (N-NN) within the same corpus. We note that the use of clean, controlled non-neutral speech samples make these EERs optimistic (Lombard effect samples would typically contain background noise).

A further consideration is that all matched and mismatched comparisons in Table II, with the exception of CHAINS N-NN, involve chunks extracted from the same recording session. While the absolute value of the resulting EERs could therefore be considered optimistic, it is the comparison of EERs across conditions that is of most interest here. Note that the CHAINS N-NN EER is slightly higher than the N-NN condition in the other whisper corpora; this performance loss can be (at least in part) attributed to recording session independence.

Given the clear performance gain with the x-vector system, the remainder of the paper will focus on scores obtained using this system only.

B. Calibration Performance

In addition to discrimination, it is important to asses shown on neutral speech may affect calibration performance. A system can be said to have good calibration performance if its scores can be used to make good decisions (using an application-specific threshold, for example). Calibration performance is dependent on the numerical values of the scores, and not just the ordering of the same-speaker and different-speaker scores (as is the case with a purely-discriminative metric, like EER). The C_{llr} (log-likelihood ratio cost) [53] measures

the quality of a set of scores for making decisions, using a logarithmic cost function. The C_{llr} is a measure of both discrimination and calibration. If the C_{llr} for a set of scores is close to 0, the system can be said to have good discrimination and calibration. If the C_{llr} for a set of scores is 1 or above, the system either has poor discrimination, poor calibration, or both. It is possible to transform the scores output by a system in order to improve their calibration performance and reduce C_{llr}; this process is referred to as score calibration. Score calibration is typically applied by shifting and scaling scores using predetermined parameters obtained from a separate set of training scores using linear logistic regression [54]. The choice of calibration data is important; for the transformation to be effective, the calibration training data must be representative of the test data being calibrated. The minimum achievable C_{llr} for a set of scores can be calculated via a monotonic transformation [53]. This C_{llr}^{min} value represents the case of 'perfect' calibration. The effectiveness of score calibration can therefore be measured by the level of mis-calibration (or calibration loss) between C_{llr} and the C_{llr}^{min}. The C_{llr}^{min} can be viewed as a theoretical minimum for a set of scores. In our subsequent experiments, we define a more practical minimum C_{llr} for measuring calibration loss.

Given the distinct neutral and non-neutral speech conditions in the Lombard effect and whisper corpora in this study, there are several score calibration scenarios to consider:

- *Neutral calibration:* linear calibration parameters (i.e. one scaling and one shifting parameter) are trained using the scores from neutral-neutral comparisons, and used to calibrate scores from all test conditions. This scenario represents a naïve system, optimized for use with neutral speech. Neutral calibration would be expected to be effective for neutral-neutral comparisons only.
- *Pooled calibration:* linear calibration parameters are trained using the scores from all comparisons, and used to calibrate scores from all conditions. This scenario represents a general-purpose system, which has had access to all conditions in advance, but is not optimized for any particular condition. Pooled calibration would be expected to be more effective than neutral calibration overall, but less effective for neutral-neutral comparisons.
- *Matched calibration:* linear calibration parameters are trained for *each* of the three possible conditions (i.e. N-N, NN-NN, and N-NN from Table II), and used to calibrate the scores from all conditions, where each test score is calibrated by the parameters from the same condition. This scenario represents a system that has access to all conditions in advance, and applies the optimal calibration parameters for each test based on ground-truth knowledge of the comparison conditions. This scenario therefore represents a theoretical upper bound for calibration performance, but one that is more realistic than C_{llr}^{min} (for this reason, we use matched calibration performance in measuring calibration loss). Matched calibration would be expected to be more effective than neutral or matched calibration overall, and will (by definition) be equally effective as neutral calibration for neutral-neutral comparisons.

In our experiments, each of these calibration scenarios was evaluated using the x-vector scores from Table II on a per-corpus basis. A leave-one-out cross-validation scheme was adopted, whereby all scores involving a particular speaker were held-out, and the remaining scores were used to learn calibration parameters for the held-out speaker. All scores involving this particular speaker were then calibrated with the trained parameters. This process was repeated for all speakers in the corpus, before calculating error metrics on the full set of calibrated scores. The cross-validation approach ensures that the training data is representative of the test data, without overlap between the scores (or the speakers) used for training and testing calibration parameters. Cross-corpus calibration would have been appropriate, but would have resulted in some loss of representativeness due to differences between the corpora (e.g., language mismatch between Pool2010 and UT-SCOPE).

Table IV provides C_{llr} values for these calibration variants applied to the x-vector comparison scores from each corpus. In addition to the performance over all test conditions (ALL), the performance breakdown for neutral-only and neutral vs. non-neutral test conditions is provided. In a similar manner to the weighted EER (Table II), the C_{llr} values for the ALL test condition are weighted to equalize the contribution of the three comparison conditions.

Referring first to the ALL test condition, it can be seen that matched calibration results in the lowest C_{llr} for each corpus, followed by pooled and then neutral calibration. This is expected, due to the condition-optimized parameters applied to each score with matched calibration. Pooled calibration has had access to all conditions, but applies only one general set of parameters to each test score. Neutral calibration has not had the benefit of non-neutral conditions in training, and is therefore performs poorly on comparisons involving non-neutral speech. It is also clear that whisper degrades calibration performance to a greater extent than Lombard effect: with matched calibration, whisper C_{llr} values for the ALL test condition are at least three times greater than their Lombard equivalent. They approximately double with pooled calibration and rise above 1 with neutral calibration. This is consistent with discrimination performance trends observed in Table II. For UT-SCOPE-LE and Lombard-Grid corpora, the difference between neutral, pooled, and matched calibration performance for the ALL test condition is relatively small, with C_{llr}s of ≈ 0.1 obtained even with neutral calibration. Within the constraints of these two corpora therefore, the Lombard effect does not cause a large drop in discrimination or calibration performance. With Pool2010 however, the C_{llr} increases by a factor of 10 from pooled to neutral (0.0620–0.669). This large calibration loss is an indicator that Pool2010 neutral samples are poorly representative of the Pool2010 non-neutral samples.

The N-N test condition breakdown in Table IV shows the C_{llr} for each calibration variant when applied to neutral-only comparisons. With neutral or matched calibration (which are equivalent in this case) the calibration performance is similar across all corpora - this is in keeping with the discrimination results in Table II. The addition of Lombard speech in the calibration training set does not greatly hurt performance, as C_{llr} increases only slightly with pooled calibration for the Lombard corpora. However, the inclusion of whisper in pooled calibration training increases the C_{llr} of neutral-only comparisons by an order of magnitude for all three corpora relative to matched calibration.

Finally, referring to the mismatched N-NN test condition of Table IV, which shows C_{llr} for each calibration variant when applied to neutral vs. non-neutral comparisons, we see the lowest C_{llr}s again with matched calibration. For Lombard corpora, there is a progressive increase in C_{llr}s from matched to pooled, and from pooled to neutral (with the exception of Pool2010, for which the C_{llr} increases dramatically with neutral calibration). A similar trend is observed with whisper corpora, but with much larger C_{llr}s. For pooled calibration, the N-NN test condition results in C_{llr}s of ≈ 1 for all whisper corpora.

The results in Table IV again show that both Lombard effect and whisper affect speaker recognition performance, with whisper having a much greater detrimental impact. Considering a scenario where a system must handle mismatched N-NN comparisons, for Lombard effect there will be significant calibration loss if only using neutral scores to train the calibration function (an approximate doubling of C_{llr} for UT-SCOPE-LE and Lombard-Grid, and a much larger increase for Pool2010). Depending on the specifics of the data, the C_{llr} may still be low in absolute terms (i.e. UT-SCOPE-LE and Lombard-Grid with C_{llr} of ≈ 0.2). For whisper, discrimination and calibration performance of mismatched N-NN comparisons is poor even with matched calibration, and with pooled or neutral calibration the C_{llr} rises to ≥1. A system provides useful information (in a forensic voice comparison case for example) when the C_{llr} is below 1; this gap between matched and pooled calibration in the case of whisper is therefore significant.

Subsequent experiments in this study propose new calibration approaches that do not require ground-truth knowledge of the speech sample conditions, and can bridge this performance gap between matched and pooled calibration.

C. Per-Speaker Performance

The measures of discrimination and calibration detailed in Tables II and IV show the performance effects of non-neutral speech on a population level. It may be the case however, that sub-groups within the population (or individual speakers), display different performance characteristics. This is particularly true for intrinsic variability, which is by definition a product of the individual speaker. Exclusively relying on population-level metrics can mask such underlying performance variations. For example, the age difference between the speakers in a non-target (i.e. different-speaker) comparison has been shown to influence speaker recognition scores[55], as has an age difference between two samples in a target (i.e. same-speaker) comparison [52]. This underlying factor of age may not be apparent from a population-level analysis.

Zoo plots are a means of visually assessing the performance of biometric systems at the per-speaker level. The concept of zoo plots was introduced by Doddington *et al.* [56], who proposed the designation of speakers as different animals based on the statistics of their comparison scores, relative to the whole population (the zoo). The zoo plot concept was subsequently extended by increasing the number of possible animal designations [57], and through the visual incorporation of score standard deviation for each of the speakers [58].

A zoo plot is created by representing each speaker as a point on a plot according to their mean same-speaker and different-speaker comparison scores. The speakers falling within the

top or bottom 25% (i.e. the first and last quartiles) on each axis are designated different animal labels. For example, the ‘best’ performing speakers, having low different-speaker scores and high same-speaker scores, are ‘doves,’ and the ‘worst’ performing speakers, having high different-speaker scores and low same-speaker scores, are ‘worms’. This designation of speakers as different animals can be a useful device for summarizing behavior of individuals or sub-groups, but we note that it is largely dependent on the system and test data [59]. Here, we are not concerned with specific animal designation, but rather with using the zoo plots as a means of visualizing individual and population level score characteristics under Lombard effect and whisper vocal effort variability.

Fig. 2 contains a combined neutral and non-neutral zoo plot for each of the corpora. Each plot contains two points (ellipses) for each speaker, one based on neutral-only scores, and the other based on neutral vs. non-neutral scores. The quartile boundaries and speaker points for the two conditions are colored in blue and red respectively. All scores have been calibrated with neutral calibration. The standard deviation of a speaker’s scores on each axis, relative to the mean of the neutral standard deviation across all speakers, is used to determine their ellipse dimensions - a speaker may be ‘tall,’ ‘short,’ ‘fat,’ or ‘thin’ [58]. The score shift that occurs between neutral and non-neutral conditions can be observed here at the individual speaker level.

With Lombard effect, at the population-level, there is a decrease in same-speaker scores (a shift to the left for all nonneutral points), and an increase in variance between speakers on the same-speaker score axis (reflected in the wider interquartile range). At the per-speaker level, some speakers move more than others in the score space across conditions, resulting in more ‘outlier’ speakers (e.g., speaker 8 in Fig. 2(e)) There is also a small increase in same-speaker score variance (width of the ellipses). These observations are shared across the three Lombard corpora.

With whisper, at the population-level, there is a decrease in same-speaker scores *and* in different-speaker scores (an upward shift to the left for all non-neutral points), and a relatively constant variance between speakers on both axes (reflected in the similar interquartile ranges). At the per-speaker level, there is a noticeable decrease in score variance on both axes (a decrease in ellipse area), and *fewer* ‘outlier’ speakers. These observations are shared across the three whisper corpora.

Inspecting these plots at the population-level, the relative score shifts for Lombard effect and whisper align with expectations, given the speaker recognition performance observed in each case. At the speaker-level, these plots suggest that between-speaker score variability increases for Lombard effect and decreases in whisper. Although beyond the scope of this study, we note that the speakers who are very mobile (or very immobile) in the zoo plot between conditions could be used to inform signal-level understanding of these sources of variability.

V. Vocal Effort Detection

The speaker recognition experiments in the previous sections have established that both Lombard effect and whisper have a detrimental impact on discrimination and calibration performance. Across different Lombard effect and whisper corpora, there are consistencies in how non-neutral speech affects the speaker recognition score space. These observations motivate the detection of Lombard effect and whisper as a first step toward addressing their detrimental impacts, and suggest that a general, corpus-independent detection approach may be feasible. While we propose a shared approach for the detection of Lombard effect and whisper, we stress that since Lombard effect and whisper occur in different scenarios, are motivated by different factors (LE is a reflex, whisper is deliberate), and have very different effects on the signal, we do not consider a three-way LE-whisper-neutral detection approach.

A. Visualizing Speaker Representations

First, to investigate the relationship between neutral and nonneutral speech in the speaker model space, x-vectors for each 10 s speech chunk were projected into two dimensions using t-Distributed Stochastic Neighbor Embedding (t-SNE) [60]. t-SNE finds a non-parametric embedding of high-dimensional data in an unsupervised way (i.e. not using class labels). Data is embedded in the t-SNE space such that similar data points appear closer together and dissimilar data points appear further apart. It is consequently a useful tool for providing visual insight into the structure of high-dimensional data. Fig. 3 provides a two dimensional t-SNE embedding of Lombard, whisper, and neutral x-vectors. The number of x-vectors plotted per-corpus was equalized by random sampling to ensure equal contribution of each corpus to the embedding transformation. We note that similar visualization can be achieved using principal components analysis (PCA) (this was applied to visualize the whisper i-vector space in our previous study [19]); however, we found t-SNE to be more visually informative with this data.

The distribution of x-vectors within the Lombard-neutral and whisper-neutral t-SNE spaces presents some interesting observations. Referring first to the Lombard-neutral space in Fig. 3(a), x-vectors are generally clustered according to corpus and speaker gender; there is no global separation of Lombard effect and neutral x-vectors. For both UT-SCOPE-LE and Lombard-Grid corpora, the small groups of points within each corpus-gender cluster consist of x-vectors for an individual speaker. For example, the UT-SCOPE-LE male cluster (red points) consists of 6 smaller groups of points, one for each of the 6 UT-SCOPE-LE male speakers. There is some separation of Lombard effect and neutral x-vectors on this per-speaker basis. For Pool2010, there is better overall separation of Lombard effect and neutral x-vectors. Pool 2010 contains only male speakers however. We note that the Pool2010 x-vectors (purple points) that fall into the same cluster as the UT-SCOPE females (darker blue points) originate from a male speaker with an unusually high fundamental frequency. Overall, this t-SNE visualization suggests that speaker gender followed by corpus characteristics (e.g., microphone and channel type, speech content, speaker language and accent) are the dominant variables within the set of Lombard-neutral x-vectors. However, the global separation of Lombard effect and neutral x-vectors for Pool2010, and per-speaker

separation for UT-SCOPE-LE and Lombard-Grid corpora suggest that a generalizable Lombard speech detector may be feasible.

Referring to the whisper-neutral space in Fig. 3(b), x-vectors are tightly clustered according to whether they correspond to neutral or whispered speech, and according to speaker gender. The gender separation is greater in neutral speech than in whisper; this is reasonable given the loss of voicing, and hence fundamental frequency, from whispered speech. This global separation of whisper and neutral across a corpus was shown for UT-SCOPE-LE and CHAINS corpora in our previous study [19]. In Fig. 3(b), points from the same corpus are generally close together, with no distinct corpus clusters. We note that the CHAINS neutral female x-vectors (yellow circles) that fall within the neutral male x-vector cluster originate from a female speaker with an unusually low fundamental frequency. Overall, this t-SNE visualization suggests that whisper, followed by speaker gender are the dominant variables within the set of whisper-neutral x-vectors. This is a strong indication that a generalizable whisper detector is feasible.

B. Non-Neutral Speech Detection Using x-vectors

The use of i-vectors as features for classification of information other than speaker identity is well established; for example, i-vectors have been successfully applied to speaker age estimation [61], spoken language recognition [62] and detection of pathological speech [63]. More recently, x-vectors have been used in a similar way for spoken language recognition [64], and to detect pathological speech [65]. In this study, we build on the proposal in [19] by evaluating a cross-corpus non-neutral speech detector based on x-vectors. In this approach, a two-class linear support vector machine (SVM) is trained using neutral and non-neutral 10 s x-vectors from one corpus, and used to classify x-vectors from a different corpus. The training x-vectors were pre-processed with mean and length normalization, and the test x-vectors were mean normalized with statistics from the training set, and then length normalized. Based on initial within-corpus cross-validation tests, a polynomial kernel was used for the Lombard effect SVM, and a linear kernel for whisper.

Fig. 4 shows the output of the cross-corpus SVM testing for each corpus: It is clear that cross-corpus detection of whisper is very effective (EER of 0%), while cross-corpus detection of Lombard effect is less accurate (EER ≈ 25%). Based on the x-vector distributions in Section V-A, this is not surprising: Lombard effect exists on a continuous scale (dependent on the noise level and type [3], [11], and on the individual speaker, Fig. 2), whereas whisper is a binary state.

In the context of speaker recognition score calibration, the detection score distributions in Fig. 4(a) motivate the inclusion of Lombard effect detection score as a continuous-valued quality measure, and whisper detection score for the selection of discrete calibration parameters. In Section VI, these score calibration proposals are evaluated.

VI. Vocal Effort Calibration

The output of the non-neutral speech detectors evaluated in Sec. V could be used to inform a speaker recognition system in multiple ways, for example:

- As a gate-keeping measure to reject non-neutral samples based on a predetermined detection score threshold.
- As a classifier to determine the conditions of a test comparison, allowing condition-specific modeling, normalization, or calibration parameters to be selected.
- As a quality measure that can be incorporated in a score-quality calibration function.

The first option involves rejecting comparisons subject to a threshold; this is important functionality for real-word systems, and an effective means of reducing calibration error [66]. Here, we consider only the second two options, which do not involve rejection of any comparisons. We note however that based on the results in Section V-B, a gatekeeping measure would clearly be effective at rejecting whisper comparisons.

The second option is considered here as an extension of matched calibration (Sec. IV-B), for which the appropriate calibration parameters for a comparison are selected based on the output of the detector rather than ground-truth knowledge of the recording conditions. This was applied to whisper comparisons in a cross-corpus manner in our previous work [19]; here we extend this approach to Lombard comparisons.

The third option is assessed here by using the detector output from the two recordings in a comparison to train a score-quality calibration function, which can be applied to new comparisons without any ground-truth knowledge of the recording conditions. A constrained version of this approach was applied in our previous work [11], in which ground-truth Lombard noise exposure levels were used to train a calibration function, requiring that ground-truth noise levels were also available at test time. Here we extend this approach by removing the need for ground-truth knowledge for test comparisons, or for a corpus-specific calibration function, and evaluate on both whisper and Lombard effect speech.

A. Score-Quality Calibration

Incorporating ‘quality’ information in score calibration using Quality Measure Functions (QMFs) is well established; QMFs have been applied to incorporate duration information [50], [67], [68], noise estimates [67], [68], and aging (i.e. time interval) information [52]. Here, we consider the use of a QMF to incorporate the likelihood of non-neutral speech as an additional term in conventional score calibration.

As discussed in Section IV-B, score calibration can be applied using a linear transformation [69], in which raw scores s are transformed into calibrated scores s' given offset and scaling parameters w_0 and w_1:

$$s' = w_0 + w_1 s, \tag{1}$$

where w_0 and w_1 are obtained by linear logistic regression optimization [54] on a separate set of representative scores. In the QMF approach, conventional linear calibration is extended by including an additional term, w_2:

$$s' = w_0 + w_1 s + w_2 Q(i), \tag{2}$$

where Q is a QMF depending on some (estimated or ground-truth) information i associated with the audio samples resulting in the scores s. The w_2 parameter governs the influence of the additional quality term. (2) can be generalized by including multiple additional terms (i.e. $w_3 \ldots w_N$), each with an associated QMF.

Here, we define information i_x as the detection score given a test sample x and a non-neutral speech detector, and i_y as the detection score given a test sample y and a non-neutral speech detector.

We consider two approaches for incorporating i_x and i_y in calibration: the first includes i_x and i_y as two independent terms, allowing the relationship between the raw scores and the non-neutral speech levels in each sample to be learned from the calibration training data. The second includes the absolute difference between i_x and i_y as a single term. From previous studies, we observe that QMFs modelling the absolute difference between i_x and i_y in this way generalize well across different datasets and measurements, including time interval [52] and duration [67]. The two score-quality calibration variants considered are:

$$Q_1 \colon\ s' = w_0 + w_1 s + w_2 i_x + w_3 i_y, \tag{3}$$

$$Q_2 \colon\ s' = w_0 + w_1 s + w_2 |i_x - i_y|, \tag{4}$$

where x and y are pairs of audio samples resulting in scores s. (3) models the non-neutral speech detection scores for samples x and y independently, while (4) models the absolute difference between detection scores for the two samples. These approaches will be referred to as Q_1 and Q_2 respectively.

B. Vocal Effort Calibration Experiments

In this Section, the two proposed score-calibration approaches are evaluated alongside several variants of conventional calibration. As in Sec. IV-B, a leave-one-out cross-validation approach is used for calibration, with non-neutral speech detection scores obtained in across-corpus fashion, as per Sec. V-B. We focus this assessment on neutral vs. non-neutral comparisons, which are the most challenging in terms of discrimination and calibration. To compare various calibration approaches, we use a measure of relative calibration loss, R_C:

$$R_C = \left(C_{llr} - C_{llr}^M\right) / C_{llr}^M, \tag{5}$$

where C_{llr}^M is the C_{llr} obtained with matched calibration. R_C therefore indicates how close calibration performance is to a *realistic* optimum, C_{llr}^M.

Table V provides R_C (%) values for several calibration approaches, for each corpus. In addition to the proposed score quality approaches Q_1 and Q_2, we include *predicted*

calibration, which is equivalent to matched calibration with non-neutral speech detector labels in place of ground-truth labels used to select the calibration parameters for a test comparison. Note that the LE detector is use for LE corpora, and the whisper detector for the whisper corpora. Pooled calibration is included as a baseline for comparison.

For Lombard effect corpora, it is clear that Q_2 calibration is the most effective approach, with R_C values lower than the pooled calibration baseline, and both Q_1 and predicted calibration. The negative R_C value for Pool2010 indicates that the C_{llr} obtained with Q_2 is lower than that of matched calibration. There is no performance improvement with Q_1 calibration or predicted calibration compared to the pooled calibration baseline. For whisper corpora, we see again that Q_2 improves on pooled calibration (and on matched calibration, in the case of CHAINS), and that Q_1 provides no improvement. Due to perfect performance of the whisper detector, predicted calibration performance is equivalent to matched performance, and therefore R_C is 0 with this approach.

With score-quality calibration functions of the form in (2), the use of per-sample quality information allows for a change in discrimination performance. The effect of the proposed Q_1, Q_2, and predicted calibration approaches on discrimination performance (in terms of EER) over *all* comparison conditions is presented in Table VI.

A similar pattern is observed to that in Table V: Q_2 generally outperforms Q_1 and the pooled baseline, and is the best choice for Lombard speech (with the exception of Lombard-Grid). The EERs remains lightly higher than those with matched calibration however. Predicted calibration is again the best performing approach with whisper, and results in the same EERs as matched calibration.

C. Lombard Effect in Noise

The data used in this study consists of exclusively clean speech, allowing for an exploration of the effects of whisper and Lombard effect without confounding variability. However, given that Lombard effect speech is induced by noise exposure, in practical applications it is likely that this noise will also be present in the audio recording (depending on the characteristics of the capture device, the level of noise present in the signal will vary; for example, directional microphones, microphone arrays, or active noise cancellation can all reduce the noise level relative to the speech level). While a full exploration of Lombard effect and noise is beyond the scope of the present paper, here we demonstrate an example of Lombard effect detection and calibration in the case where the noise inducing Lombard effect is present in the signal.

For this experiment we selected a constrained set of UT-SCOPE-LE data, consisting of Lombard effect samples induced by large crowd noise at 80 dB-SPL, along with the neutral samples from the same speakers. A noisy version of this set was generated by adding large crowd noise to each sample. Based on the noise source level of 80 dB-SPL, and the close proximity of the microphone to the speaker in the UT-SCOPE-LE collection, a value of 0 dB was chosen as a suitably representative SNR [70]. Speaker recognition performance was then assessed for three conditions: clean neutral vs. clean LE speech, clean neutral vs. noisy LE speech, and noisy neutral vs. noisy LE speech.

For Lombard effect detection in this experiment, we considered the inclusion of noisy samples into the model training set. A noisy version of the LombardGrid data (which, like UT-SCOPE-LE, contains LE data induced by 80 dB-SPL noise) was created by adding random selections of background noise from the MUSAN dataset [71], at SNRs of between 0 and 10 dB SNR, to all samples. Along with the Lombard effect detector introduced in Section V-B (trained with clean speech only), we trained two further Lombard effect detectors, the first with clean neutral and noisy LE speech, and the second with noisy neutral and noisy LE speech.

Table VII presents the calibration performance of the calibration approaches proposed in Section VI-B for each of the three new Lombard effect conditions. In addition to Q_1 and Q_2 calibration, which use the output of an Lombard effect detector trained with clean speech, we evaluate Q_1^M and Q_2^M, which use the output of an Lombard effect detector matched ('M') to the noise condition of the comparison (e.g. for the noisy neutral vs. noisy LE condition, 'No-No,' the Lombard effect detector is trained with noisy neutral and noisy LE samples). We omit predicted calibration performance, as it is outperformed by all of the quality-based calibration approaches.

Referring to Table VII, the drastic effect of additive noise at 0 dB SNR is evident from the increase in $C_{llr}^{\min}$ and C_{llr}^{M} from the 'Cln-Cln' condition to both of the noisy conditions. With pooled calibration, the calibration loss in terms of R_{C}% is particularly high in the 'Cln-No' condition (54.43%), where there is the largest mismatch in terms of noise. Q_2 calibration is very effective in the 'Cln-Cln' condition, with 0% calibration loss relative to matched calibration. Q_1 and Q_2 are ineffective in the noisy conditions, which is not unexpected, given their exclusively clean training data. If noise is introduced into the detector training however, the Q_1^M and Q_2^M calibration performance greatly improves, outperforming pooled calibration and approaching that of matched calibration.

Finally, in Table VIII we present discrimination performance for all comparison conditions within the clean/noisy Lombard effect data; we again see here the large performance drop due to additive noise, and that with Q_1^M and Q_2^M calibration, performance approaches that achieved with matched calibration. The improvement offered by Q_2^M calibration relative ot pooled calibration is particularly noticeable in the 'Cln-No' condition, where the EER drops from 20.41% to 11.66%.

VII. Conclusions

This study has presented a series of controlled speaker recognition experiments involving Lombard effect (LE) and whisper, two commonly-occurring modes of non-neutral speech. In line with expectations, both speech modes had a detrimental impact on speaker recognition discrimination and calibration performance relative to neutral-only comparisons, with greater impact in the presence of whisper. Through the use of zoo plot and t-SNE visualizations, there were consistencies observed in population-level and speaker-level behavior in the presence of LE and whisper. These observations motivated the use of non-neutral speech detection to inform score calibration, which led to an improvement in

performance across all corpora. A constrained experiment involving LE with additive noise demonstrated that with a suitably trained LE detector, this approach is also effective in noisy conditions.

The relatively low EERs and C_{llr}s obtained for neutral vs. LE comparisons suggest that Lombard effect need not be an obstacle in the application of speaker recognition, if the scores are calibrated appropriately. Lombard effect 'in the wild' presents a greater challenge, due to the likely presence of the noise stimulus in the comparison samples. Our constrained experiments with Lombard effect in noise suggest that discrimination will suffer, but some calibration loss is recoverable using the proposed approaches.

Despite the performance drop with neutral vs. whisper comparisons, C_{llr}s of less than 1 were obtained with the proposed calibration approaches, demonstrating that neutral vs. whisper speaker recognition has potential for use in certain contexts: for example, to add evidential value in a forensic voice comparison case. Depending on the application, it may be more practical to reject comparisons involving whisper, or switch to a whisper-specific system (with modifications at the front-end, for example). The accuracy observed with the whisper detector here indicate that such a gate-keeping mechanism would be a viable option.

The framework proposed in this study - SVM-based nonneutral speech detection with simple quality measure functions for score calibration - has the potential to be applied to other forms of intrinsic speaker variability beyond Lombard effect and whisper. Specifically, the approach most successful with Lombard effect, which used the absolute difference between Lombard effect detection scores in calibration (i.e. Q_2 calibration), could be extended to other vocal effort variations existing on a continuous scale. For example, speech produced under different levels of 'situational' stress, either cognitive or physical [5], [72] [34]. In our previous study [52], aging information was integrated into calibration using the absolute time difference between the recording dates of the two samples in the comparison (a similar approach to Q_2 calibration in this study). We note that the Q_1 calibration approach, which considered the non-neutral speech detection scores from the two samples in a comparison independently, would likely benefit from additional training data.

The approach most successful with whisper was to use the output of the detector to select predetermined calibration parameters; this could potentially be applied to other forms of intrinsic variability that typically exist in a binary state, for example: falsetto speech, non-speech vocalizations such as laughter or scream [73] (in such extreme cases, the best option is likely to be rejection of the comparison), or language switching.

Overcoming intrinsic variability is of key importance for unconstrained applications of speaker recognition, such as those encountered in forensic and investigative domains. Ensuring appropriate score calibration in the presence of intrinsic variability enables the use of speaker recognition to contribute evidence in the form of a likelihood ratio, or to make reliable decisions to inform an investigation. The approach presented in this study demonstrates the use of a non-neutral speech detector to achieve calibration performance close to that obtained using ground-truth non-neutral speech labels. This framework can be

applied independently of the speaker recognition system, and has potential to be extended to other forms of intrinsic speaker variability.

Acknowledgments

This work was supported by the University of Texas at Dallas from Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen, and in part by grants from BCOE 09-097MM-UTD, CTTSO PO22450, and AFRL FA8750-15-1-0205. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Alberto Abad.

Biography

Finnian Kelly received the B.A.I. engineering and Ph.D. degrees from Trinity College Dublin, Dublin, Ireland. He is currently with The University of Texas at Dallas, Richardson, TX, USA, where he joined the Center for Robust Speech Systems as a Research Associate in October 2014, and in January 2016, he joined Oxford Wave Research Ltd., as a Research Scientist. He is a Member of the Research Committee of the International Association for Forensic Phonetics and Acoustics and an Affiliate Member of the NIST OSAC Speaker Recognition subcommittee. His research interests include audio, speech, and speaker analysis, and the use of automatic speaker recognition in forensic applications.

John H. L. Hansen (Fellow, IEEE) received the B.S.E.E. degree from the College of Engineering, Rutgers University, New Brunswick, NJ, USA, and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, GA, USA. In 2005, he joined the Erik Jonsson School of Engineering and Computer Science, the University of Texas at Dallas, Richardson, TX, USA, where he is currently an Associate Dean for research and a Professor of electrical and computer engineering. He also holds the Distinguished University Chair in telecommunications engineering and a joint appointment as a Professor of speech and hearing with the School of Behavioral and Brain Sciences. From 2005 to 2012, he was the Head of the Department of Electrical Engineering, the University of Texas at Dallas. At UT Dallas, he established the Center for Robust Speech Systems. From 1998 to 2005, he was the Department Chair and a Professor of speech, language, and hearing sciences, and a Professor of electrical and computer engineering with the University of Colorado Boulder, Boulder, CO, USA, where he co-founded and was an Associate Director of the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory. He has supervised 92 Ph.D. or M.S. thesis students, which include 51 Ph.D. and 41 M.S. or M.A. He has authored

or coauthored 765 journal and conference papers including 13 textbooks in the field of speech processing and language technology, signal processing for vehicle systems, co-author of the textbook *Discrete-Time Processing of Speech Signals* (IEEE Press, 2000), *Vehicles, Drivers and Safety: Intelligent Vehicles and Transportation* (vol. 2 DeGruyter, 2020), *Digital Signal Processing for In-Vehicle Systems and Safety* (Springer, 2012), and the lead author of The Impact of Speech Under 'Stress' on Military Speech Technology (NATO RTO-TR-10, 2000). His research interests include machine learning for speech and language processing, speech processing, analysis, and modeling of speech and speaker traits, speech enhancement, signal processing for hearing impaired or cochlear implants, machine learning-based knowledge estimation and extraction of naturalistic audio, and in-vehicle driver modeling and distraction assessment for human-machine interaction. He is an IEEE Fellow for contributions to robust speech recognition in stress and noise, and ISCA Fellow for contributions to research for speech processing of signals under adverse conditions. He was the recipient of Acoustical Society of America's 25 Year Award in 2010, and is currently serving as ISCA President (2017–2022). He is also a Member and the past Vice-Chair on U.S. Office of Scientific Advisory Committees (OSAC) for OSAC-Speaker in the voice forensics domain from 2015 to 2021. He was the IEEE Technical Committee (TC) Chair and a Member of the IEEE Signal Processing Society: Speech-Language Processing Technical Committee (SLTC) from 2005 to 2008 and from 2010 to 2014, elected the IEEE SLTC Chairman from 2011 to 2013, and elected an ISCA Distinguished Lecturer from 2011 to 2012. He was a Member of the IEEE Signal Processing Society Educational Technical Committee from 2005 to 2010, a Technical Advisor to the U.S. Delegate for NATO (IST/TG-01), an IEEE Signal Processing Society Distinguished Lecturer from 2005 to 2006, an Associate Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING from 1992 to 1999 and the IEEE SIGNAL PROCESSING LETTERS from 1998 to 2000, Editorial Board Member for the IEEE *Signal Processing Magazine* from 2001 to 2003, and the Guest Editorin October 1994 for Special Issue on Robust Speech Recognition for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. He is currently an Associate Editor for the JASA, and was on the Speech Communications Technical Committee for Acoustical Society of America from 2000 to 2003. In 2016, he was awarded the honorary degree Doctor Technices Honoris Causa from Aalborg University, Aalborg, Denmark in recognition of his contributions to the field of speech signal processing and speech or language or hearing sciences. He was the recipient of the 2020 Provost's Award for Excellence in Graduate Student Supervision from the University of Texas at Dallas and the 2005 University of Colorado Teacher Recognition Award. He organized and was General Chair for ISCA Interspeech-2002, Co-Organizer and Technical Program Chair for the IEEE ICASSP-2010, Dallas, TX, and Co-Chair and Organizer for IEEE SLT-2014, Lake Tahoe, NV. He will be the Tech. Program Chair for the IEEE ICASSP-2024, and Co-Chair and Organizer for ISCA INTERSPEECH-2022.

References

[1]. Greenberg CS, Mason LP, Sadjadi SO, and Reynolds DA, "Two decades of speaker recognition evaluation at the national institute of standards and technology," Comput. Speech Lang, vol. 60, 2020, Art. no. 101032.

[2]. Lombard E, “Le signe de l’elevation de la voix,” Ann. Mal. De L’Oreille et du larynx, vol. 37, pp. 101–119, 1911.

[3]. Hansen JHL and Varadarajan V, “Analysis and compensation of lombard speech across noise type and levels with application to in-set/outof-set speaker recognition,” IEEE Trans, Audio, Speech, Lang. Proc, vol. 17, no. 2, pp. 336–378, Feb. 2009.

[4]. Bořil H. and Hansen JHL, “UT-Scope: Towards LVCS Runder Lombard effect induced by varying types and levels of noisy background,” in Proc. Int. Conf. Acoust., Speech, Signal Process, 2011, pp. 4472–4475.

[5]. Hansen JHL, “Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition,” Speech Commun, vol. 20, no. 2, pp. 151–173, Nov. 1996.

[6]. Hansen JHL et al. , “The impact of speech under ‘stress’ on military speech technology,” NATO Res. Tech. Organization RTO-TR-10, AC/323(IST)TP/5IST/TG-01, Mar. 2000.

[7]. Bořil H. and Hansen JHL, “Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments,” IEEE Trans. Audio, Speech, Lang. Proc., vol. 18, no. 6, pp. 1379–1393, Aug. 2010.

[8]. Junqua JC, “The Lombard reflex and its role on human listeners and automatic speech recognizers,” J. Acoust. Soc. Amer, vol. 93, no. 1, pp. 510–524, 1993. [PubMed: 8423266]

[9]. Chi S-M and Oh Y-H, “Lombard effect compensation and noise suppression for noisy Lombard speech recognition,” in Proc. Int. Conf. Spoken Lang. Process., 1996, pp. 2013–2016.

[10]. Saleem MM, Liu G, and Hansen JHL, “Weighted training for speech under Lombard effect for speaker recognition,” in Proc. Int. Conf. Acoust., Speech, Signal Process, 2015, pp. 4350–4354.

[11]. Kelly F. and Hansen JHL, “Evaluation and calibration of Lombard effects in speaker verification,” in Proc. IEEE Spoken Lang. Technol.Workshop, San Diego, CA, USA, 2016, pp. 205–209.

[12]. Masthoff H, “A report on a voice disguise experiment,” Int. J. Speech, Lang., Law, vol. 3, no. 1, pp. 160–167, 1996.

[13]. Ghaffarzadegan S, Boril H, and Hansen JHL, “UT-VOCAL EFFORT II: Analysis and constrained-lexicon recognition of whispered speech,” in Proc. Int. Conf. Acoust., Speech, Signal Process. 2014, pp. 2544–2548.

[14]. Ito T,Takeda K,andI.F,“Analysisandrecognitionofwhisperedspeech,” Speech Commun, vol. 45, no. 2, pp. 139–152, 2005.

[15]. Fan X. and Hansen JHL, “Speaker identification within whispered speech audio streams,” IEEE Trans, Audio, Speech, Lang. Proc., vol. 19, no. 5, pp. 1408–1421, Jul. 2011.

[16]. Vestman V,Gowda D,Sahidullah M,Alku P,and Kinnunen T,“Speaker recognition from whispered speech: Atutorial survey and an application of time-varyinglinearprediction,”SpeechCommun,vol.99,no.1,pp.62–79, 2018.

[17]. Sarria-Paja M. and Falk TH, “Fusion of auditory inspired amplitude modulation spectrum and cepstral features for whispered and normal speechspeakerverification,”Comput.Speech,Lang,vol.45,pp.437–456, 2017.

[18]. Das RK and Li H, “On the importance of vocal tract constriction for speaker characterization: The whispered speech study,” in Proc. Int. Conf. Acoust., Speech, Signal Process, 2020, pp. 7119–7123.

[19]. Kelly F. and Hansen JHL, “Detection and calibration of whisper for speaker recognition,” in Proc. IEEE Spoken Lang. Technol. Workshop, Athens, Greece, 2018, pp. 1060–1065.

[20]. Brümmer N. and du Preez J, “Application-independent evaluation of speaker detection,” Comput. Speech, Lang, vol. 20, no. 2–3, pp. 230–275, 2006.

[21]. Nandwana MK, McLaren M, Ferrer L, Castan D, and Lawson A, “Analysis and mitigation of vocal effort variations in speaker recognition,” in Proc. Int. Conf. Acoust., Speech, Signal Process, 2019, pp. 6001–6005.

[22]. McLaren M, Lawson A, Ferrer L, Scheffer N, and Lei Y, “Trial-based calibration for speaker recognition in unseen conditions,” Proc. Odyssey, 2014, pp. 19–25.

[23]. Nandwana MK, Ferrer L, McLaren M, Castan D, and Lawson A, “Analysis of critical metadata factors for the calibration of speaker recognition systems,” Proc. INTERSPEECH, 2019, pp. 4325–4329.

[24]. Ferrer L. and McLaren M, “A speaker verification backend for improved calibration performance across varying conditions,” 2020, arXiv:2002.03802.

[25]. Dehak N, Kenny P, Dehak R, Dumouchel P, and Ouellet P, “Front-end factor analysis for speaker verification,” IEEE Trans. Audio, Speech, Lang. Process, vol. 19, no. 4, pp. 788–798, May 2011.

[26]. Snyder D, Garcia-Romero D, Sell G, Povey D, and Khudanpur S, “xvectors: Robust DNN embeddings for speaker recognition,” in Proc. Int. Conf. Acoust., Speech, Signal Process, 2018, pp. 5329–5333.

[27]. Ikeno A, Varadarajan V, Patil S, and Hansen JHL, “UT-Scope: Speech under Lombard effect and cognitive stress,” in Proc. IEEE Aerosp. Conf, 2007, pp. 1–7.

[28]. Jessen M,Koster O,andS.Gfroerer,“Influenceofvocaleffortonaverage and variability of fundamental frequency,” Int. J. Speech, Lang., Law, vol. 2, no. 12, pp. 174–21, 2005.

[29]. Alghamdi N, Maddock S, Marxer R, Barker J, and Brown GJ, “A corpus of audio-visual Lombard speech with frontal and profile views,” J. Acoust. Soc. Amer, vol. 143, no. 6, pp. 523–529, 2018.

[30]. Zhang C. and Hansen JHL, “Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing,” IEEE Trans. Audio, Speech, Lang. Proc., vol. 19, no. 4, pp. 883–894, May 2011.

[31]. Cummins F, Grimaldi M, Leonard T, and Simko J, “The chains corpus: characterizing individual speakers,” in Proc. SPECOM, 2006, pp. 431–435.

[32]. Lim BP, “Computational differences between whispered and non-whispered speech,” Ph.D. dissertation, Dept. Elect. Comput. Eng., Univ. Illinois at Urbana-Champaign, 2010.

[33]. Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS, and Dahlgren NL, TIMIT Acoustic-phonetic Continuous Speech Corpus LDC93S1. Philadelphia: Linguistic Data Consortium, 1993. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S1

[34]. Zhang C, Liu G, yu C, and Hansen JHL, “I-Vector Based physical task stress detection with different fusion strategies,” INTERSPEECH, Sep. 2015, pp. 2689–2693.

[35]. Boril H, Grezl F, and Hansen JHL, “Front-end compensation methods for LVCSR under Lombard effect,” INTERSPEECH, 2011, pp. 1257–1260.

[36]. Enzinger E. and Kasess CH, “Bayesian vocal tract model estimates of nasal stops for speaker verification,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process, 2014, pp. 1685–1689.

[37]. Jessen M, Alexander A, and Forth O, “Forensic voice comparisons in German with phonetic and automatic features using VOCALISE software,” in Proc. Audio Eng. Soc. Forensics Conf, 2014, pp. 1–8.

[38]. Kirchhuebel C, “The effects of Lombard speech on vowel formant measurements,” J. Acoust. Soc. Amer, vol. 128, no. 4, p. 2395, 2010.

[39]. Marxer R, Barker J, Alghamdi N, and Maddock S, “The impact of the Lombard effect on audio and visual speech recognition systems,” Speech Commun, vol. 100, pp. 56–68, 2018.

[40]. Michelsanti D, Tan Z-H, Sigurdsson S, and Jensen J, “Deep-learning-based audio-visual speech enhancement in presence of Lombard effect,” Speech Commun, vol. 115, pp. 38–50, 2019.

[41]. Fan X. and Hansen JHL, “Acoustic analysis and feature transformation from neutral to whisper for speaker identification within whispered speech audio streams,” Speech Commun, vol. 55, pp. 119–134, 2013.

[42]. Ghaffarzadegan S, Boril H, and Hansen JHL, “Deep neural network training for whispered speech recognition using small databases and generative model sampling,” Int. J. Speech Tech, vol. 20, pp. 1063–1075, 2017.

[43]. Grimaldi M. and Cummins F, “Speaker identification using instantaneous frequencies,” IEEE Trans. Audio, Speech, Lang. Process, vol. 16, pp. 1097–1111, Sep. 2008.

[44]. Sarria-Paja M, Senoussaoui M, O’Shaughnessy D, and Falk T, “Feature mapping, score-, and feature-level fusion for improved normal and whispered speech speaker verification,” in Proc. Int. Conf. Acoust., Speech, Signal Process, 2016, pp. 5480–5484.

[45]. Kenny P, “Bayesian speaker verification with heavy-tailed priors,” Odyssey, 2010, Paper 14.

[46]. Sadjadi SO and Hansen JHL, “Unsupervised speech activity detection using voicing measures and perceptual spectral flux,” IEEE Signal Process. Lett, vol. 20, no. 3, pp. 197–200, Mar. 2013.

[47]. Kinnunen T. and Li H, “An overview of text-independent speaker recognition: From features to supervectors,” Speech Commun, vol. 52, no. 1, pp. 12–40, 2010.

[48]. Garcia-Romero D. and Espy-Wilson CY, “Analysis of i-vector length normalization in speaker recognition systems,” INTERSPEECH, 2011, pp. 249–252.

[49]. Hasan T, Saeidi R, Hansen JHL, and van Leeuwen DA, “Duration mismatch compensation for i-vector based speaker recognition systems,” in Proc. Int. Conf. Acoust., Speech, Signal Process., Vancouver, Canada, 2013, pp. 7663–7667.

[50]. Mandasari M, Saeidi R, McLaren M, and van Leeuwen D, “Quality measure functions for calibration of speaker recognition system in various duration conditions,” IEEE Trans. Audio, Speech, Lang. Process, vol. 21, no. 11, pp. 2425–2438, Nov. 2013.

[51]. Leeuwen D. A. v., “A note on performance metrics for speaker recognition using multiple conditions in an evaluation,” Tech. Rep, 9 Jun. 2008. [Online]. Available: https://www.researchgate.net/publication/237138418_A_note_on_performance_metrics_for_Speaker_Recognition_using_multiple_conditions_in_an_evaluation

[52]. Kelly F. and Hansen JHL, “Score-aging calibration for speaker verification,” IEEE/ACM Trans. Audio, Speech, Lang. Proc., vol. 24, no. 12, pp. 2414–2424, Dec. 2016.

[53]. D. A. v. Leeuwen and N. Brümmer, “An introduction to application-independent evaluation of speaker recognition systems,” Speaker Classification, vol. 1, pp. 330–353, 2007.

[54]. Pigeon S, Druyts P, and Verlinde P, “Applying logistic regression to the fusion of the NIST ‘99 1-speaker submissions,” Digit. Signal Process, vol. 10, no. 1–3, pp. 237–248, 2000.

[55]. Doddington G, “The effect of target/non-target age difference on speaker recognition performance,” Odyssey, 2012, pp. 263–267.

[56]. Doddington G, Liggett W, Martin AF, Przybocki MA, and Reynolds DA, “Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation,” in Proc. Int. Conf. Spoken Lang. Process., 1998, Paper 608.

[57]. Yager N. and Dunstone T, Biometric Systems for Data Analysis: Design, Evaluation, and Data Mining. Berlin, Germany: Springer Press, 2009.

[58]. Alexander A,Forth O,Nash J,and Yager N,“Zooplotsforspeakerrecognition with tall and fat animals,” in Proc. Int. Assoc. Forensic Phonetics Acoust. Conf, 2014, pp. 1–2.

[59]. Teli MN, Beveridge JR, Phillips PJ, Givens GH, Bolme DS, and Draper BA, “Biometric zoos: Theory and experimental evidence,” in Proc. Int. Joint Conf. Biometrics, 2011, pp. 1–8.

[60]. van der Maaten LJP and Hinton GE, “Visualizing high-dimensional data using t-SNE,” J. Mach. Learn. Res, vol. 9, pp. 2579–2605, 2008.

[61]. Bahari MH, McLaren M, Van hamme H, and van Leeuwen DA, “Speaker age estimation using i-vectors,” Eng. Appl. Artif. Intell, vol. 34, no. C, pp. 99–108, 2014.

[62]. Dehak N, Torres-Carrasquillo P, Reynolds D, and Dehak R, “Language recognition via i-vectors and dimensionality reduction,” Proc. INTERSPEECH, Jan. 2011, pp. 857–860.

[63]. García N, Vásquez-Correa JC, Orozco-Arroyave JR, and Nöth E, “Multimodal i-vectors to detect and evaluate parkinson’s disease,” INTERSPEECH, 2018, pp. 2349–2353.

[64]. Snyder D, Garcia-Romero D, McCree A, Sell G, Povey D, and Khudanpur S, “Spoken language recognition using x-vectors,” Odyssey, 2018, pp. 105–111.

[65]. Botelho C,Teixeira F,Rolland T,Abad A,and Trancoso I,“Pathological speechdetectionusingX-vectorembeddings,”2020,arXiv:2003.00864v2.

[66]. Ferrer L, Nandwana MK, McLaren M, Castán D, and Lawson A, “Towardfail-safe speaker recognition: Trial-based calibration with a reject option,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 27, no. 1, pp. 140–153, Jan. 2019.

[67]. Mandasari MI, Saeidi R, and van Leeuwen DA, “Quality measures based calibration with duration and noise dependency for speaker recognition,” Speech Commun, vol. 72, pp. 126–137, 2015.

[68]. Nautsch A, Saeidi R, Rathgeb C, and Busch C, “Robustness of quality-based score calibration of speaker recognition systems with respect to low-SNR and short-duration conditions,” Odyssey, 2016, pp. 358–365.

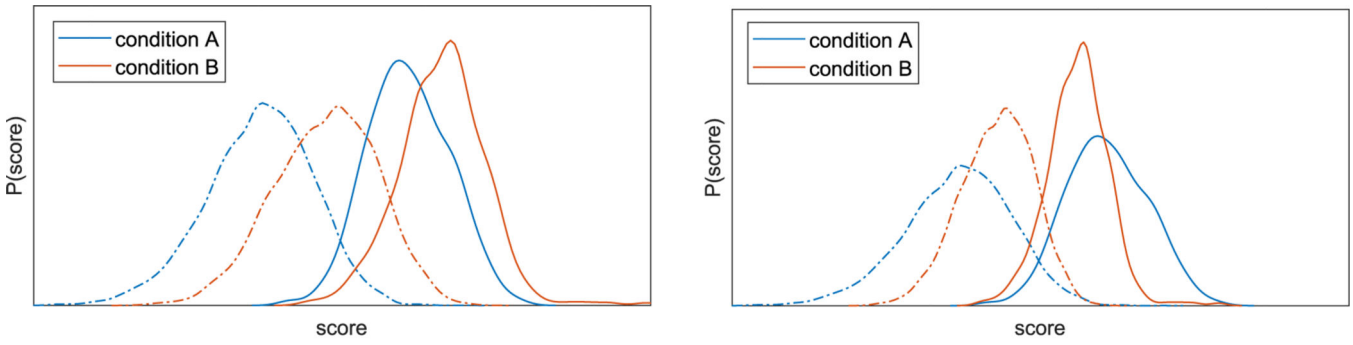
[69]. Brümmer N, van Leeuwen DA, and Swart A, “A comparison of linear and non-linear calibrations for speaker recognition,” Odyssey, 2014, pp. 14–48.

[70]. Weisser A. and Buchholz JM, “Conversational speech levels and signal to-noise ratios in realistic acoustic conditions,” J. Acoust. Soc. Amer, vol. 145, pp. 249–360, 2019.

[71]. Snyder D, Chen G, and Povey D, “Musan: A music, speech, and noise corpus,” 2015, arXiv:1510.08484v1.

[72]. Godin K. and Hansen JHL, “Physical task stress and speaker variability in voice quality,” EURASIP J. Audio, Speech, Music Process, vol. 2015, no. 29, pp. 1–13, 2015.

[73]. Hansen JHL, Nandwana MK, and Shokouhi N, “Analysis of human scream and its impact on text-independent speaker verification,” J. Acoust. Soc. Amer, vol. 141, no. 4, pp. 2957–2967, 2017. [PubMed: 28464689]



(a) Poorly-calibrated scores (b) Well-calibrated scores

Fig. 1.

Distributions of same-speaker scores (solid lines) and different-speaker scores (dashed lines) for comparisons in different conditions, A and B, with a poorly-calibrated system and a well-calibrated system.

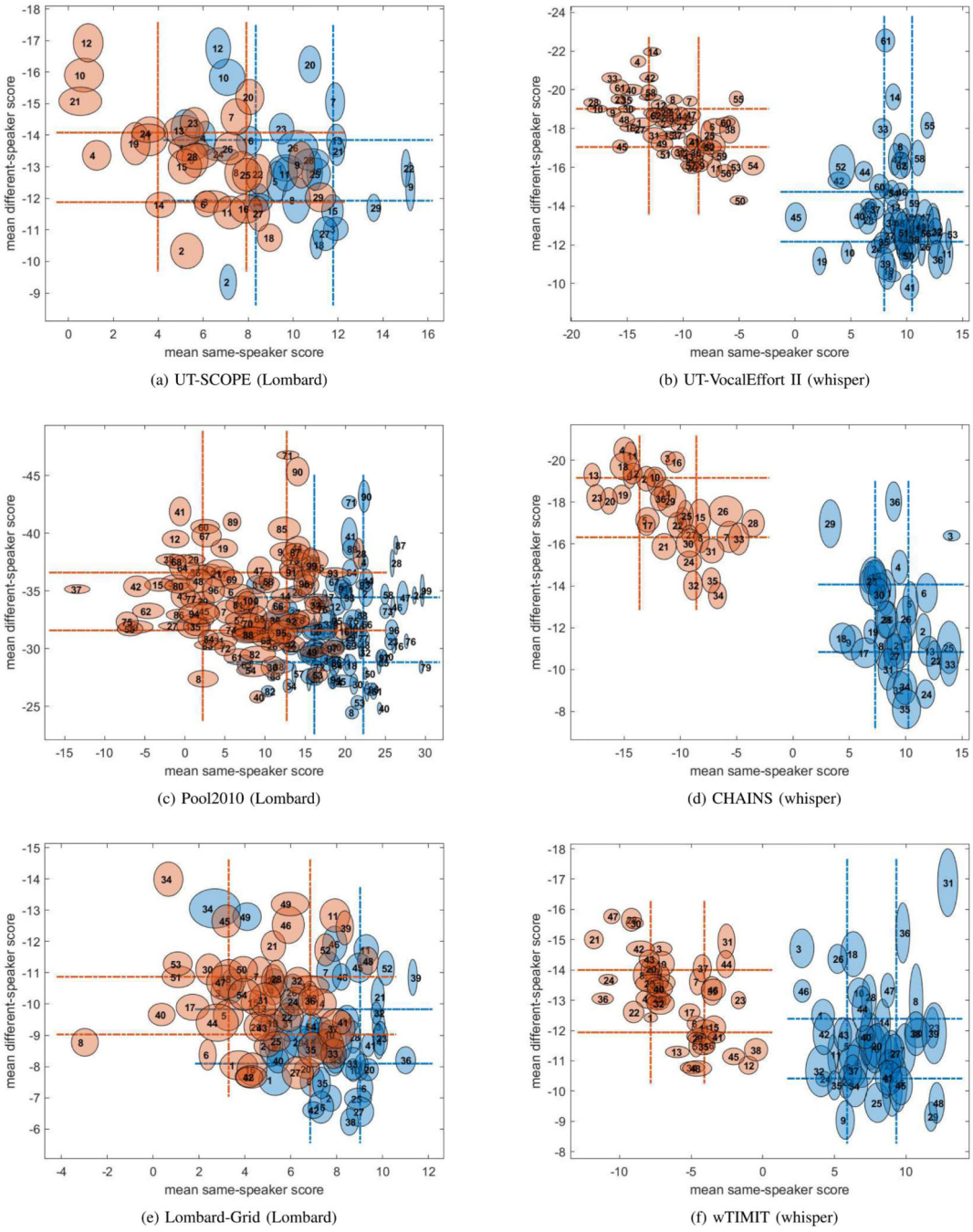
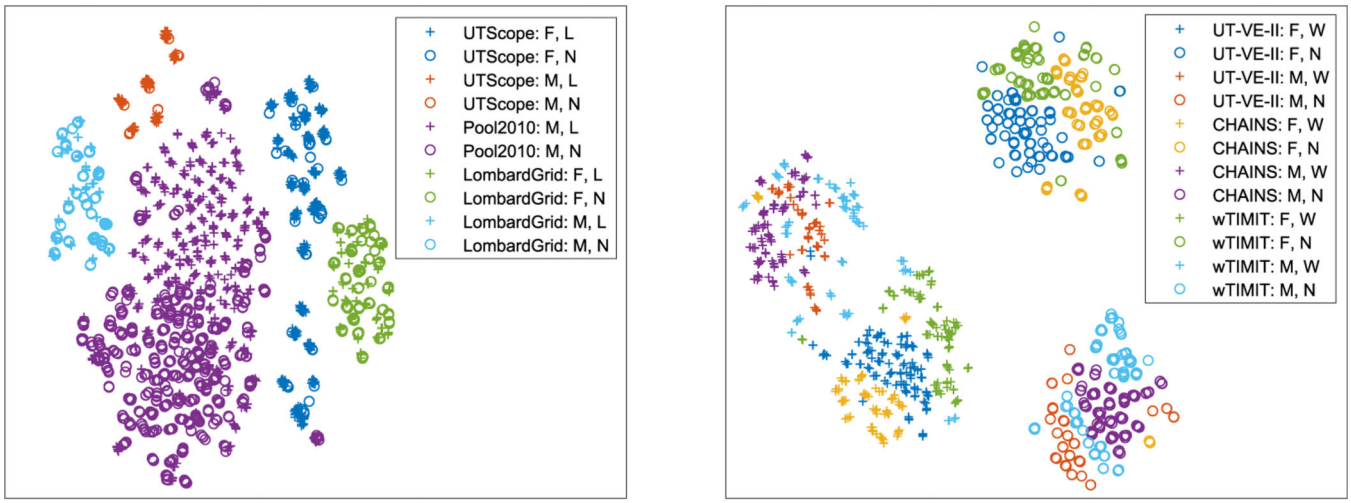


Fig. 2.

Zoo plots showing the movement in per-speaker scores between neutral vs. neutral (blue) comparisons and neutral vs. non-neutral (red) comparisons. Neutral calibration has been applied to all scores using a leave-one-out cross-validation approach on a per-corpus basis. Each (uniquely numbered) speaker is represented by one ellipse each condition, with its dimensions determined by the score variability.

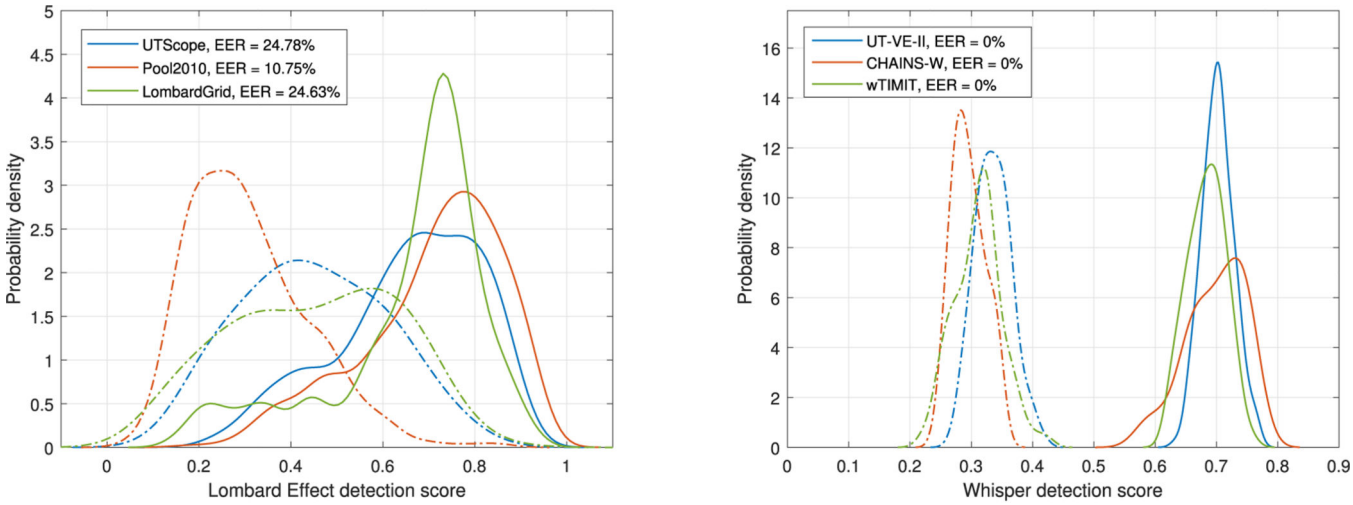


(a) Neutral and Lombard effect x-vectors

(b) Neutral and whisper x-vectors

Fig. 3.

t-SNE embedding of x-vectors for neutral and non-neutral 10 s speech chunks. 3(a) shows the x-vectors for Lombard speech (L, denoted by +) and neutral speech (N, denoted by **o**) across the three Lombard corpora. 3(b) shows the x-vectors for whisper (W, denoted by +) and neutral speech (N, denoted by **o**) speech across the three whisper corpora. For each corpus, female (F) and male (M) speakers are represented by different colors (Pool2010 contains only male speakers).



(a) Lombard effect detection

(b) Whisper detection

Fig. 4.

Non-neutral speech detection: SVM output score distributions and associated EERs. In each plot, solid lines indicate the score distribution for non-neutral test x-vectors, and dashed lines indicate the score distribution for neutral test x-vectors. All SVM training and testing is cross-corpus: for Lombard, UT-SCOPE is used to train the SVM for testing Lombard-Grid, and Lombard-Grid is used to train the SVM for testing both UT-SCOPE and Pool2010. For whisper, UT-VE-II is used to train the SVM for testing CHAINS, and CHAINS is used to train the SVM for testing both UT-VE-II and wTIMIT.

TABLE I

SPEECH CORPORA SUMMARY

corpus	speech modes	# speakers (m/f)†	language (accent)	recording location	total speech†	used in
UT-SCOPE-LE [27]	Lombard, neutral	30 (6/24)	English (US)	studio	11.6 hrs	[3] [4] [34] [35]
Pool2010 [28]	Lombard, neutral	100 (100/0)	German	studio	2.4 hrs	[36] [37] [38]
Lombard-Grid [29]	Lombard, neutral	54 (24/30)	English (UK)	studio	3.8 hrs	[39] [40]
UT-VE II [30]	whisper, neutral	62 (20/42)	English (US)	studio	1.5 hrs	[15][41][13][42][19]
CHAINS [31]	whisper, neutral	36 (20/16)	English (Irish, US, UK)	studio, office	7.1 hrs	[43][19]
wTIMrr [32]	whisper, neutral	48 (24/24)	English (US, Singapore)	studio	58.4 hrs	[44][17]

Summary of the speech corpora content used in this study. All recordings consist of read speech recorded with a close microphone, with samples truncated to a fixed duration of net-speech.

†Note that several of these corpora contain content not used in this study, including recordings of additional speakers, recordings of spontaneous speech, and recordings with alternative microphones.

IEEE/ACM Trans Audio Speech Lang Process. Author manuscript; available in PMC 2022 June 30.

TABLE II

DISCRIMINATION PERFORMANCE WITHIN EACH CORPUS

	N-N	NN-NN	N-NN	ALL
i-vector				
UT-SCOPE (LE)	5.63	9.07	10.78	8.98
Pool2010 (LE)	2.12	1.43	10.08	5.60
Lombard-Grid (LE)	3.33	3.10	8.26	5.74
UT-VE II (W)	8.23	10.86	30.12	28.47
CHAINS (W)	4.29	10.13	30.95	27.57
wTIMIT (W)	7.62	13.56	26.74	25.26
x-vector				
UT-SCOPE (LE)	0.41	1.82	2.54	2.11
Pool2010 (LE)	0.02	0.03	2.60	1.52
Lombard-Grid (LE)	1.73	2.12	3.62	2.94
UT-VE II (W)	0.85	6.06	19.88	24.72
CHAINS (W)	0.95	5.96	23.74	26.85
wTIMIT (W)	1.67	7.36	19.56	24.58

EERs (%) for matched neutral vs. neutral (N-N) and non-neutral vs. non-neutral (NN-NN) comparisons, and mismatched neutral vs. non-neutral (N-NN) comparisons, and the pooled set of comparisons across these three conditions (ALL), for each lombard effect (LE) and whisper (W) corpus. All comparisons involved 10 second speech chunks, and the number of trials for each condition is shown in Table III.

TABLE III

THE NUMBER OF TRIALS FOR ALL CONDITIONS AND CORPORA

	N-N	NN-NN	N-NN	ALL
# target trials				
UT-SCOPE (LE)	370	33236	7695	41301
Pool2010 (LE)	262	355	896	1513
Lombard-Grid (LE)	540	450	1350	2430
UT-VE II (W)	360	360	900	1620
CHAINS (W)	480	480	1200	2160
wTIMIT (W)	960	960	2400	4320
# non-target trials				
UT-SCOPE (LE)	16900	1203550	285450	1505900
Pool2010 (LE)	124868	100096	175652	400616
Lombard-Grid (LE)	35550	35550	71100	142200
UT-VE II (W)	29104	28352	57480	114936
CHAINS (W)	15500	15500	31000	62000
wTEVQT (W)	27600	27600	55200	110400

The number of target trials (same-speaker comparisons) and non-target trials (different-speaker comparisons) for each condition in each corpus.

TABLE IV

CALIBRATION PERFORMANCE WITHIN EACH CORPUS

Train:	Neutral	Pooled	Matched
Test:		ALL	
UT-SCOPE (LE)	0.112	0.083	0.073
Pool2010 (LE)	0.669	0.062	0.031
Lombard-Grid (LE)	0.134	0.113	0.103
UT-VE II (W)	3.025	0.743	0.3
CHAINS (W)	3.330	0.773	0.329
wTIMIT (W)	1.950	0.676	0.213
Test:		N-N	
UT-SCOPE (LE)	0.037	0.040	0.037
Pool2010 (LE)	0.003	0.033	0.003
Lombard-Grid (LE)	0.076	0.087	0.076
UT-VE II (W)	0.039	0.366	0.039
CHAINS (W)	0.049	0.441	0.049
wTTMIT (W)	0.067	0.311	0.067
Test:		N-NN	
UT-SCOPE (LE)	0.222	0.125	0.106
Pool2010 (LE)	2.000	0.117	0.088
Lombard-Grid (LE)	0.244	0.168	0.154
UT-VE II (W)	8.026	1.022	0.614
CHAINS (W)	7.915	1.046	0.718
wTIMIT (W)	4.45	0.944	0.597

C_{llr} for 10 second x-vector comparisons with three calibration variants: neutral, pooled, and matched. Three test comparison conditions are considered: neutral vs. neutral (N-N), neutral vs. non-neutral (N-NN), and the pooled set of scores from all conditions (ALL). Each of the calibration variants is applied with a leave-one-out cross-validation approach on a per-corpus basis.

TABLE V

SCORE-QUALITY CALIBRATION PERFORMANCE

			R_C%			
	C_{llr}^{min}	C_{llr}^{M}	**pooled**	Q_1	Q_2	**pred.**
UT-SCO. (LE)	0.105	0.106	17.92	22.64	**4.72**	25.47
Pool2010 (LE)	0.082	0.088	32.95	35.23	**−7.95**	132.95
L.-Grid (LE)	0.142	0.151	11.26	11.92	**9.93**	15.89
UT-VE II (W)	0.603	0.614	66.45	77.69	17.43	**0.0**
CHAINS (W)	0.705	0.718	45.68	45.54	**−12.67**	0.0
wTIMIT (W)	0.586	0.597	61.31	61.31	14.57	**0.0**

Neutral vs. non-neutral 10 second x-vector comparisons: C_{llr}^{min}, matched calibration performance (C_{llr}^{M}), and calibration loss relative to matched calibration (R_C%) for: pooled calibration, Q_1 calibration (uses the non-neutral speech detection scores of both comparison samples independently, (3)), Q_2 calibration (uses the absolute difference between non-neutral speech detection scores as a quality measure, (4)), and predicted calibration (uses the non-neutral speech detection scores to select matched calibration parameters). Note that negative percentages indicate a *lower* C_{llr} than matched calibration.

TABLE VI

SCORE-QUALITY DISCRIMINATION PERFORMANCE

EER%					
	matched	**pooled**	Q_1	Q_2	**predicted**
UT-SCOPE (LE)	1.66	2.16	2.17	**2.01**	2.20
P00I2OIO (LE)	0.78	1.65	1.65	**0.88**	1.29
Lombard-Grid (LE)	2.58	**2.95**	2.90	2.97	3.13
UT-VE II (W)	9.16	24.77	25.76	15.10	**9.16**
CHAINS (W)	10.53	26.91	25.08	10.78	**10.63**
wTIMIT (W)	9.69	24.58	23.87	13.92	**9.69**

10 second x-vector comparisons for all conditions: discrimination performance (EER%) with matched calibration, pooled calibration, Q_1 calibration (uses the non-neutral speech detection scores of both comparison samples independently), Q_2 Calibration (uses the absolute difference between non-neutral speech detection scores as a quality measure), and predicted calibration (uses the non-neutral speech detection scores to select matched parameters).

TABLE VII

SCORE-QUALITY CALIBATION PERFORMANCE IN NOISE

			R_C%				
	$C_{llr}^{\min}$	C_{llr}^{M}	**pool.**	Q_1	Q_2	Q_1^M	Q_2^M
Cln-Cln	0.077	0.105	25.00	27.38	**0.00**	27.38	**0.00**
Cln-No	0.555	0.564	54.43	53.55	54.96	52.84	**6.91**
No-No	0.657	0.682	5.41	5.41	9.43	**3.55**	10.66

10 second x-vector comparisons of UT-SCOPE neutral vs le samples induced by crowd noise at 80 dB-SPL, with and without additive crowd noise at 0 dB SNR. Only neutral vs LE comparisons are considered: 875 target trials, 32 550 non-target trials. ‘cln-Cln’ denotes clean neutral vs. clean LE, ‘cln-No’ denotes clean neutral vs noisy LE, and ‘no-No’ denotes noisy neutral vs. noisy LE. The calibration approaches and R_C metric are the same as in Table V, with the addition of matched (‘M’) condition Q_1^M and Q_2^M calibration.

TABLE VIII

SCORE-QUALITY DISCRIMINATION PERFORMANCE IN NOISE

EER%						
	matched	**pooled**	Q_1	Q_2	Q_1^M	Q_2^M
Cln-Cln	1.33	1.95	1.83	**1.62**	1.83	**1.62**
Cln-No	10.72	20.41	20.28	20.51	17.84	**11.66**
No-No	17.78	18.34	18.48	18.41	**17.72**	18.02

10 second x-vector comparisons of UT-SCOPE neutral and LE samples induced by crowd noise at 80 dB-SPL, with and without additive crowd noise at 0 dB SNR. All comparison conditions are considered: 1605 target trials, 65 100 non-target trials. ‘cln-Cln’ denotes clean neutral and clean LE, ‘cln-No’ denotes clean neutral and noisy LE, and ‘no-No’ denotes noisy neutral and noisy LE samples. The calibration approaches are the same as in Table V, with the addition of matched (‘M’) condition Q_1^M and Q_2^M calibration.