On the Limits of Automatic Speaker Verification: Explaining Degraded Recognizer Scores Through Acoustic Changes Resulting from Voice Disguise

Rosa González Hautamäki,[1, a)] Ville Hautamäki,[1] and Tomi Kinnunen[1]
School of Computing, University of Eastern Finland, Joensuu, P-O-BOX 111, 80110, Finland[b]

In *speaker verification* research, objective performance benchmarking of listeners and automatic speaker verification (ASV) systems is of key importance in understanding the limits of speaker recognition. While the adoption of common data and metrics has been instrumental to progress in ASV, they have two major shortcomings. First, the utterances lack intentional voice changes imposed by the speaker. Second, the standard evaluation metrics focus on average performance across all speakers and trials. As a result, a knowledge gap remains in how the acoustic changes impact recognition performance at the level of individual speakers. We address the limits of speaker recognition in ASV systems under voice disguise using a *linear mixed effects model* to analyze the impact of change in long-term statistics of selected features (formants F1-F4, their bandwidths B1-B4, F0 and speaking rate) to ASV log-likelihood ratio (LLR) score. The correlations between the proposed predictive model and the LLR scores are 0.72 for females and 0.81 for male speakers. As a whole, the difference in long-term F0 between enrollment and test utterances was found to be the individually most detrimental factor, even if the ASV system uses only spectral, rather than prosodic, features.

©*2019 Acoustical Society of America.* [http://dx.doi.org(DOI number)]

[XYZ] Pages: 1–14

I. INTRODUCTION

The task of *speaker recognition* — recognizing persons from their voices (Hansen and Hasan, 2015; Schmidt-Nielsen and Stern, 1985) — can be performed by listeners and automatic systems. A major source of performance degradation of automatic speaker verification (ASV) systems is condition mismatch between the test and the reference, or enrollment, utterances. The standard datasets used in the field have been primarily designed to address performance factors related to channel and environment (Doddington *et al.*, 2000; Lei and Hansen, 2016), text-dependency (Garofolo *et al.*, 1993; Larcher *et al.*, 2014), and duration (Lee *et al.*, 2015) to name a few. Interestingly, however, performance degradation due to *within-speaker* variation has received far less attention even though it has a strong impact on the accuracy of ASV systems (Kahn *et al.*, 2010). Within-speaker variation arises from the speaker and can include changes in pronunciation, speaking style, short-term health condition, emotion, or vocal effort. Vocal effort was addressed in one of the National Institute of Standards Technology (NIST) speaker recognition evaluation (SRE) campaigns (Greenberg *et al.*, 2011) involving the *Lombard reflex*, which refers to the automatic raising of one's voice under noisy environments for the purpose of increasing intelligibility.

Even if most ASV systems utilize methods to normalize within-speaker variation — *within-class covariance normalization* (Hatch *et al.*, 2006) and *probabilistic linear discriminant analysis* (PLDA) (Prince and Elder, 2007)) being examples — they have two major limitations. First, because of training data limitations, they model combined variations in speaker, channel, content, duration, and other factors; these variations are lumped into combined *session variation*, making it difficult to disentangle speaker-related and other effects. Further, as noted by Ajili (2017), evaluation corpora are often more focused on inter-speaker effects and contain relatively few, or too homogeneous, recordings of the same target speaker. The second limitation of most prior studies is the *extent* of within-speaker variation that is considered; ASV systems are rarely subjected to real stress tests involving *extreme* within-speaker variation, in specific, variations that are intentionally introduced to avoid detection. Indeed, many ASV studies make the implicit assumption that the speaker is either *cooperative* (wants to be recognized as him or herself) or is *unaware* of being subjected to ASV testing. The former holds in authentication applications (such as online banking) and the latter in screening and search applications, such as the monitoring of telephone calls and targeted voice search from the Internet. Under the 'cooperative or unaware' situations, the speaker is less likely to *deliberately* modify his or her voice (Hansen and Hasan, 2015; Rodman and Powell, 2000). Changes in vocal effort and short-term

[a)] Electronic mail: rgonza@cs.uef.fi, villeh@cs.uef.fi, tkinnu@cs.uef.fi
[b] The following article has been accepted by the Journal of the Acoustical Society of America. After it is published, it will be found at http://asa.scitation.org/journal/jas

health conditions are examples of *non-deliberate* variation.

Because of these shortcomings, there is currently no detailed picture of the detrimental effects of *deliberate* within-speaker voice modifications to ASV accuracy. Voice acting, disguise and mimicry (González Hautamäki *et al.*, 2016; Leemann and Kolly, 2015; Zhang, 2012) are examples of deliberate voice modifications that can substantially impact ASV accuracy. We focus on *disguise*, the act of purposeful attempt to conceal one's identity. As a motivating quantitative example, in our recent study (González Hautamäki *et al.*, 2016) the equal error rate (EER) of a standard i-vector PLDA system was increased from 5.1 % to 31.7 % on high-quality clean speech data when everything else was held constant but normal (collaborative) test utterances were replaced with disguised versions involving deliberate 'age' modification. Observing such dramatic ASV performance degradations even under highly idealized conditions suggests that ASV systems are potentially far more sensitive to purposefully enforced within-speaker variation than one may expect. Detrimental effects of other forms of within-speaker variation and disguise have been reported in a number of independent studies other that our work (reviewed in Section II), in addition to effects due to physical constrictions such as handkerchiefs (Zhang and Tan, 2008) and face garments (Saeidi *et al.*, 2016). These studies have demonstrated ASV performance degradations under a variety of conditions but few have focused on *explaining* it in terms of acoustic within-speaker variation resulting from disguise; this is precisely our aim. We argue that understanding the cause of such performance degradation can be helpful to improve ASV technology and understand its limits.

The primary aim of ASV research is to improve *predictors* — namely, classifiers that predict whether to accept an identity claim based on a test utterance (with *a priori* unknown speaker identity) and a reference (enrollment) utterance with a known speaker identity. We instead aim to *explain* the behavior of a given ASV system on evaluation data with known ground-truth speaker labels of all utterances. Our study expands the methodology toolbox of ASV evaluation towards an *interpretative* framework beyond the commonly used evaluation metrics (such as *equal error rate*) that provide no further insight beyond a numerical summary. As illustrated in Fig. 1, we present methodology to explain ASV system behavior in terms of within-speaker acoustic variation. In particular, we want to relate variation in long-term statistics of interpretable segmental (formants and their bandwidths) and prosodic (F0 and speaking rate) features to the variation in the ASV system's output score. The ASV, a data-driven system constructed through machine learning techniques, is effectively treated as a 'black box'. Thus, in principle, the methodology can be applied to analyze the behavior of any ASV system.

The methodology selected to explain ASV score variation in terms of acoustic variation uses the *linear mixed effects model* (Bates *et al.*, 2015). Linear mixed effects models are a class of powerful statistical techniques to model *grouped* data with a dependency structure. These models have been around for quite a while via the development of the *restricted maximum likelihood* (REML) technique (Patterson and Thompson, 1971) but have received thus far little attention within the speech technology field.

II. RELATED WORK

Our work resides in the broad landscape of *within-speaker style variation* study with a focus on ASV performance degradation. Humans are highly flexible at adapting their speaking style to the needs of a given communication environment. Examples of spontaneous speech style variation due to *conversational telephone speech* and *interview speech* are addressed in the context of NIST SREs data and style variation across different TV (Ajili, 2017) and internet-video (Chung *et al.*, 2018) interviews crawled from YouTube. Additionally, the impact of *vocal effort* ranging from *whisper* (Vestman *et al.*, 2018) to *shout* (Hanilci *et al.*, 2013) and *scream* (similar to shout but lacking phonemic structure) (Hansen *et al.*, 2017) has been addressed in many studies. Other examples include *acted* speech by naive or professional (Pietrowicz *et al.*, 2017) speakers, *pet-directed speech* (Park *et al.*, 2018), and the impact of varied speech rate (Dellwo *et al.*, 2015). Finally, the *voice disguise* effect has been addressed in terms of faked foreign accents (Leemann and Kolly, 2015), raised or lowered pitch (Zhang, 2012), and relative age category modification (González Hautamäki *et al.*, 2017). Disguises defined in this way are no different from acted speech: the goal is to sound different from one's own voice.

Even if the individual studies cannot easily be compared due to the adoption of different corpora, experimental protocols, methods, and performance metrics, the consensus is that within-speaker style variation has a profound impact on ASV performance. However, it is not easy to relate changes in acoustic features to changes in ASV performance. Most of the prior work focuses on *either* the detrimental effect of ASV system resulting from speech style change *or* changes in acoustic features. The novelty of this work relies in linking these two through linear mixed effects models.

III. RESEARCH DATA AND STUDY AIMS

A. AVOID corpus

Our research data consists of a corpus of modal and disguised speech that is called AVOID (*Age-related Voice Disguise*) (González Hautamäki *et al.*, 2018). The corpus, now publicly available (González Hautamäki *et al.*, 2018), contains 60 speakers: 31 females and 29 males between 18 to 73 years. The read material consists of 13 sentences (11 Finnish and 2 English) recorded on two sessions for a total of 78 sentences per speaker (González Hautamäki *et al.*, 2017, Appendix A). The speakers were

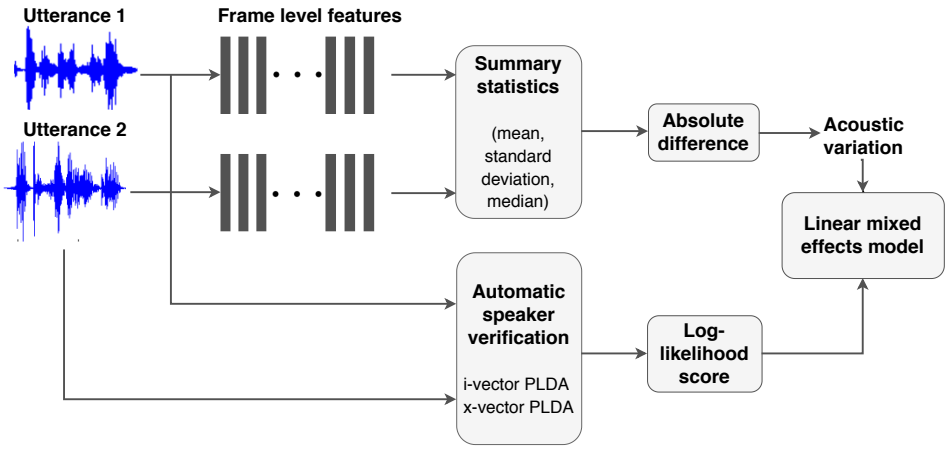


FIG. 1. Procedure for the acoustical comparison of two utterances and assessment with log-likelihood ratios (LLR). The long term features are extracted from the trial utterances and then represented by summarized statistics. The features are compared and the absolute difference is estimated to represent the acoustic variation of the trial. This variation is then used to model the LLR score associated with the trial. Automatic speaker verification (ASV) systems in this study include the i-vector and x-vector, both with probabilistic linear discriminant analysis (PLDA) scoring back-end.

TABLE I. Our mixed effect models uses a total of 15 predictor features, formed from the following combinations of features and their long-term statistical summary measures.

Acoustic features, f		Summary statistic, $\varphi(\cdot)$
Formant frequencies and bandwidths [Hz]	F1 to F4 B1 to B4	mean
Fundamental frequency [Hz]	F0	mean std. deviation median mode min max
Speech rate [syllables/s]	sr	—

instructed to read the texts both in their *modal* voice and in two disguise styles, *intended old* and *intended child*. These style modifications are common to all speakers, but are flexible enough to allow speakers to interpret *their* age-related stereotypes. Therefore, AVOID is a dataset of *acted speech* produced by *naive* speakers. According to post-hoc perceptual evaluation (González Hautamäki *et al.*, 2018) not all the produced age stereotypes are convincing to listeners but they are sufficient in increasing ASV error rates substantially, despite idealized clean data conditions.

B. Methodology - the high-level view

In our previous work (González Hautamäki *et al.*, 2018, 2017), we used the AVOID corpus to analyze the degradation of ASV (and listener) accuracy due to disguise. When a speaker is enrolled (trained) with a modal voice and tested with either of the two disguised voices, the ASV system score lowers substantially in contrast to modal-modal comparison and causes the ASV system to reject the speaker with a high probability (González Hautamäki, 2017).The idea of the present study is to provide *explanation* for the target speaker LLR score in terms of acoustic variation between the enrollment and test utterances. To this end, a given ASV system, g, is treated as a black-box measurement device that outputs the LLR score (*response variable*, y) between the two speech utterances U_1 and U_2,

$$y = g(U_1, U_2|\boldsymbol{\theta}_{\text{asv}}), \tag{1}$$

represented by parameters $\boldsymbol{\theta}_{\text{asv}}$ that encapsulate all the acoustic front-end and data-driven components of the ASV system. The higher the value of y, the more confident the ASV system is that the speaker identities of U_1 and U_2 agree. In addition, we extract *acoustic distance* (*predictor variable*, x) between U_1 and U_2, as

$$x = |\varphi(f(U_1)) - \varphi(f(U_2))|, \tag{2}$$

where $f(\cdot)$ denotes a short-term feature extractor to convert an utterance into a sequence of scalar features, while $\varphi(\cdot)$ denotes a fixed summary statistics operator. For example, if f is an F0 extractor, and φ is the sample mean, x is the distance of the average F0 values in U_1 and U_2. The lower the x, the more acoustically similar the two utterances are.

By including several features $f(\cdot)$ with meaningful combinations of $\varphi(\cdot)$, we obtain a total of D acoustic distances $x_1, \ldots, x_D$ for any pair of utterances. Our selected features are summarized in Table I and details are presented in Section VI.

Methodology-wise, our approach bears resemblance to *quality-based score calibration* such as in the work of Mandasari *et al.* (2015). Quality-based calibration adjusts the 'default' speaker similarity score produced by an ASV system towards specific operating conditions with the help of statistics such as signal-to-noise ratio (SNR) or log duration; these statistics are computed from the test and/or the enrollment utterances. These statistics are typically called quality features. Besides the *predictive vs. explanatory* aspect, our work is differentiated from quality calibration models in terms of the statistical model. The quality-calibration models assume that the trial-list of scores are independent- and identically distributed. This assumption is clearly violated because the same speaker appears multiple times as an enrollment speaker or a test-speaker, so there is a group structure defined by the speaker as a random variable in the trial list. If the group structure is not modeled explicitly, it will cause the model to *underfit*. If the task is to predict as accurately as possible, as in quality calibration literature, the underfit will not lead to wrong results, although it will not lead to the best possible performance. However, if the task is inferential as is the case here, the conclusions can be unreliable unless the group structure is accounted for (Bates *et al.*, 2015).

C. Study aims and outlook

In brief, this study aims at exploring the impact of changes in acoustical features to the performance of ASV systems scores considering the variations due to the speaker effect. The remainder of this paper is organized as follows. Section IV provides an overview of the mixed effects model. We provide the details of our ASV systems (to obtain y) in Section V and the details of predictor variables (to obtain x) in Section VI. The results are represented in two sections. First, Section VII examines the acoustic features individually, and analyzes the variation due to voice modifications. Second, Section VIII presents the results of mixed effect modeling. Specifically, we begin with an analysis of the effect of voice condition (modal and disguise voices) on variable y (LLR scores). We then explore the acoustic feature differences as predictors by ranking them based on the explanatory information that they add to the model. The model 'goodness' is evaluated in terms of the correlation between the fitted values and the modeled variable. Inferring the model from acoustical differences, individually and in groups, aims to learn the speaker-dependent and speaker-independent variability contained in the evaluation of speaker verification.

IV. LINEAR MIXED EFFECTS MODEL

The *linear mixed effects model* (Bates *et al.*, 2015) is a class of regression techniques used to model *grouped data*. Here, the ASV system's output scores LLRs form our observations, and we explain them in terms of acoustic and prosodic within-speaker variations. The general idea in *regression models* is to model a *response* y or *dependent* variable, with the focus on how other known variables, *predictors* $\boldsymbol{x} = (1, x_1, \ldots, x_D)$, explain the variation of y. The goal in the estimation is to find the values for the parameters (coefficients) $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_D)$. The coefficient β_0 is known as the *intercept* or *bias*. The coefficients of each predictor are unknown fixed constants that are common to all observations. As none of these variables are stochastic, they are known as *fixed effects*. When we add a stochastic predictor, we arrive at a *random intercept linear mixed effects model* (Bates *et al.*, 2015)

$$y_{ij} = \boldsymbol{\beta}^t \boldsymbol{x}_{ij} + b_i + \epsilon_{ij}, \tag{3}$$

where the dependent variable, y_{ij}, is the ASV score for trial j and speaker i, $\boldsymbol{\beta}^t \boldsymbol{x}_{ij}$ is the fixed effect part, b_i is the per-speaker *random effect* and ϵ_{ij} is the residual variation. The assumption for a random speaker effect and the residual error is that they are independent of each other and follow a normal distribution:

$$b_i \sim \mathcal{N}(0, {\sigma_b}^2)$$

and

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2).$$

Note that by setting $b_i = 0, \forall i$, Eq. (3) reduces to the classical linear regression model as a special case.

The purpose of modeling the response variable is to estimate the importance of each predictor from the observed data. For example, the sign of the weight β_d reveals whether the effect of predictor x_d is to *decrease* or *increase* the value of the dependent variable. In the case of studying *same speaker* or *genuine* trials, as done here, a decrease in the LLR score means that it is easier to misclassify such a trial as a different speaker trial. The larger the decrease, the larger the negative effect on the ASV system's performance.

The *residual error* models variations that are not explained by the predictors and it is defined as the difference between the response variable and the expected fitted value. An additional part of linear mixed effect models is the *random* effect, b_i. This corresponds to an effect of repeated measurements that can be correlated because they belong to the *group* factor. Assuming there is more than one trial (observation) per speaker, measurements from different trials are grouped per speaker to model between-speaker variance.

For our data, the voice condition is considered as an explanatory variable or *fixed effect* across the speakers and spoken utterances. We assume that there will be different ASV scores for different voice conditions from the same speaker. The model can reflect these individual differences by assuming different random intercepts for each speaker. In addition to by-speaker variation, we expect a 'random' variation between different sentences uttered by the same speaker.

V. AUTOMATIC SPEAKER VERIFICATION SYSTEMS

We model the LLR scores of two ASV systems based on different speaker embedding methods, namely *i-vectors* (Dehak *et al.*, 2011) and *x-vectors* (Snyder *et al.*, 2018) to demonstrate the generality of our interpretative framework. At the time of writing, the i-vector can be said to have reached a *de facto* status as a methodology that is widely-adopted by the research community, while the x-vector is a promising emerging method that is representative of the current research trends and leverages from deep neural networks.

A. I-vector and x-vector systems

Both ASV systems use mel-frequency cepstral coefficients (MFCCs) as input features and PLDA back-ends (Prince and Elder, 2007) for speaker similarity scoring. The main difference between the types of embeddings is in unsupervised generative training through Gaussian mixture modeling (i-vector) vs. speaker-discriminative training through deep neural networks (x-vector) and adoption of a longer temporal context in the x-vector system through a time-delay neural network model. Implementation-wise, the two systems are unrelated: the i-vector system is our in-house implementation that is used in our previous studies, while the x-vector system that is based on the Kaldi toolkit (Povey *et al.*, 2011) is public-domain code.

For the i-vector system, a 54-dimensional feature vector consisting of 18 MFCCs are extracted from 30 ms Hamming windowed frames. The Δ and Δ^2 features are appended to RASTA-filtered MFCCs. No speech activity detector (SAD) is used. Gender-dependent universal background models (UBMs) of 512 diagonal covariance Gaussians are trained using the *expectation-maximization* (EM) algorithm (Dempster *et al.*, 1977). A simplified PLDA with 200-dimensional speaker subspace is used in scoring.

The x-vector system uses 24 MFCCs extracted from 25 ms frames that are mean-normalized over a sliding window of three seconds. Δ features are not included. An energy-based SAD is used to discard non-speech frames. The system implementation is the Kaldi recipe with NIST SRE 16 development data models for LDA and PLDA.

B. Performance of the ASV systems

Before proceeding with our mixed effects modeling, we report the accuracy of our ASV systems on the standard protocol of the AVOID corpus in Table II. The results are presented in terms of *equal error rate* (EER), which corresponds to equal miss and false alarm rate.

The x-vector system outperformed the i-vector system, as expected. Importantly, however, neither system was immune to the disguised voice conditions. For the i-vector PLDA system, EER increases four- and six-fold for females and six- and nine-fold for males for the intended old and the intended child conditions respectively. The x-vector system performance, though more accurate in the modal voice condition, experiences *larger* relative degradation under disguise: 10-fold for females and 10- to 16-fold for males. These degradations caused by intentional voice modifications of the speaker prompts us to analyze the distribution of LLR scores to find a model that could explain the within-speaker variability.

TABLE II. ASV performance in terms of EER (%) per gender and voice condition of the test utterance. Target speakers are enrolled with modal voice and tested with three different voice conditions.

	Female		Male	
Voice condition	i-vector	x-vector	i-vector	x-vector
Modal	5.68	1.73	2.96	1.59
Intended old	24.23	17.12	18.78	15.12
Intended child	32.90	17.62	29.06	25.46

C. Dependent variable: Same speaker scores

The model concerns *genuine* (same speaker) trials only, which are those scores resulting from pairwise utterance comparisons with matched speaker identities. The LLR scores are calibrated using logistic regression with a target prior of 0.5 and false alarm and miss costs set to 1. Focal toolkit (Brümmer *et al.*, 2007) was used for the calibration, which was trained using the modal voice data and applied to the scores of the three conditions. Further analysis consider the scores pooled over the voice conditions.

VI. PREDICTOR VARIABLES: ACOUSTIC AND PROSODIC FEATURES

Among the possible alternatives to be used as predictor variables, we focus on short-term and prosodic features that are used in speaker characterization studies. We require the features to be easy to extract automatically because the AVOID corpus lacks phone-level transcription. Formants and their bandwidths are a natural choice for the short-term features for two reasons. First, they may reflect articulatory changes across the modal and disguised utterances of a given speaker. Second, change in the short-term spectral envelope (as used by the ASV systems) impacts formants, and we therefore expect them to form reasonable predictors of the LLR score. Concerning prosody, there is a vast body of literature ranging from frame-level F0 characterization (Mary and Yegnanarayana, 2008), stylized (Adami, 2007; Shriberg *et al.*, 2005) and polynomially modeled F0 contours (Dehak *et al.*, 2007), along with energy and timing- or rhythm-related features (Ajili *et al.*, 2018; Dellwo *et al.*, 2012). A number of studies have addressed the impact of such parameters in a forensic context (Lee-

mann *et al.*, 2014; Moez *et al.*, 2016). In this study, we focus on two important prosodic features: characterization of the frame-level F0 (*i.e.* without modeling their temporal envelopes), and speech rate. Both are motivated by noting the ease by which they are altered even by naive actors. Further, the dependency of the short-term spectral envelope on F0 (in particular, high F0) is known (El-Jaroudi and Makhoul, 1991), making F0-related features another potentially strong predictor of ASV system's LLR score.

The raw acoustic measurements — formants F1 to F4 and F0 — are the same as those in our earlier study (González Hautamäki *et al.*, 2017), while speech rate and formant bandwidths were added for this study. All our selected features are *frequency* measurements, so that the absolute difference operator $|\cdot|$ to produce a distance measure in the same measurement unit (Hz) has a transparent meaning.

A. Formant frequencies and bandwidths

We extract the first four formants, F1 to F4, and their bandwidths, B1 to B4, with the aid of the Burg algorithm (Childers, 1978) in Praat (Boersma and Weenink, 2015). We extract the formants from the full utterance using a window frame length of 15 ms. The maximum formant frequency was set at 5 kHz. Raw formant measurements are known to be susceptible to a number of measurement errors, such as due to a breathy voice or high F0. In our estimations, we noted many of the formant distributions to be bi-modal, particularly in the disguised voices. In our earlier study (González Hautamäki *et al.*, 2017, Appendix B), we devised a bi-Gaussian model fitted to raw F1-F3 measurements (F4 was retained as-is), which is adopted in this study as well. The mean of the lower Gaussian was selected as the long term representative formant mean of the utterance. The formants bandwidth, B1 to B4, per utterance are represented by their mean value without further processing.

B. Fundamental frequency

We used the autocorrelation-based F0 tracker (Boersma, 1993) in Praat (Boersma and Weenink, 2015) to extract F0 for each utterance every 15 ms. We used gender-specific F0 range settings: $[75, 400]$ Hz for male and $[100, 600]$ for female. Due to the high pitched 'child' voices, we set the F0 search ranges to be wider than one would typically apply for modal speech.

Previously, we used the average F0 over all frames as a scalar summary of a specific utterance and studied the relative change in the average F0 values between normal and disguised variants of the same utterance spoken by the same speaker (González Hautamäki *et al.*, 2017). In this study, we revise our F0 analysis in two respects. First, we quantify the extent of possible F0 doubling and halving errors with the aid of a *log-normal tied mixture model* (Sönmez *et al.*, 1997). The model fits a three-mode mixture distribution to the log-F0 values using an *expectation-maximization* (EM) algorithm (Dempster *et al.*, 1977). In this model, the lowest and highest modes are assumed to correspond to F0 halving and doubling errors respectively, while the middle mode represents the correct F0 values. See Fig. 2 for an example of the original F0 distributions by voice condition. We use the mean of the middle mode to select the raw F0 values that will be the feature vector for the utterance. Table III depicts the mean percentages for the F0 halving, doubling, and correct values per gender and voice condition.

TABLE III. Mean percentages of F0 halving, F0 doubling, and selected F0 presented by gender and voice condition.

		Halved	Doubled	F0
Female	Modal	0.32	0.24	99.44
	Intended old	0.68	0.34	98.98
	Intended child	0.41	0.10	99.49
Male	Modal	0.002	0.15	99.85
	Intended old	0.08	0.25	99.67
	Intended child	0.29	0.07	99.64

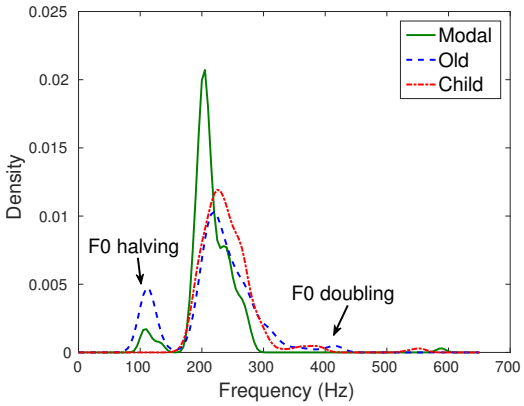


FIG. 2. (Color online) An illustration of the F0 distribution with log-tied-mix-model fit. This figure depicts the distribution with F0 halving and doubling for the female speakers in the three voice conditions (modal, intended old, and intended child) for utterance S07 (in Finnish): "Pohjantuuli ja aurinko väittelivät, kummalla olisi enemmän voimaa, kun he samalla näkivät kulkijan, jolla oli yllään lämmin takki" (in English: The North Wind and the sun were disputing which was the stronger when a traveller came along wrapped in a warm cloak.)

Second, we adopt the following simple statistics besides the average value: median, standard deviation, mode, maximum, minimum. The median is similar to the mean but is less sensitive to outliers, such as those that are potentially caused by halving or doubling errors (Farrús *et al.*, 2007). Similar to the statistical mean and median, the mode brings important information about

the F0 feature. Even if the mode is the same as the mean and median for normally distributed data, it may be very different in highly skewed distributions, which we may expect in extreme variations of F0 for certain speakers. Minimum and maximum F0 values are analyzed to consider the extreme and lowest bounds that a speaker reaches in certain utterances for the intended voice. The summary statistics were computed from the halving/doubling-compensated values.

C. Speech rate

We measure speech rate using a PRAAT implementation (De Jong and Wempe, 2009) that automatically estimates the speech rate (number of syllables / total time) by detecting syllable nuclei (Wang and Narayanan, 2007) and pause duration. The algorithm considers all peaks above a certain threshold (median intensity of the speech file) as possible syllable nuclei. To discard peaks within each syllable, the intensity measure of the utterance is used to discard consecutive peaks that do not differ by at least 2 dB in intensity. Peaks that correspond to unvoiced segments according to pitch contour are discarded. In this way, speech rate is calculated automatically without the need for transcriptions.

VII. RESULTS: MEASURING FEATURE VARIATIONS

The acoustic differences between the trial utterances are represented by differences of the summary statistics between enrollment and test utterances of each speaker. We visualize the differences in F0 statistics as boxplots in Figures 3 A and B. As expected, the mean, median and mode of F0 have similar variations. The standard deviation (SD) and minimum F0 indicate less variation, while the maximum F0 indicates the widest variation, especially for the male speakers. A one-way analysis of variance (ANOVA) (Casella and Berger, 2002) was conducted to compare the effect of the voice condition and the acoustic features summary statistics differences. A significant effect of voice condition and features differences was found with $p < 0.001$. Post-hoc comparison using *Tukey's honest significant difference (HSD)* test at 95 % significance level indicates greater differences between modal and intended child voices. An effect size $\eta^2 = 12$ % can be considered a medium size, based on Cohen's guidelines (Cohen, 1988).

Differences in speech rate were also inspected and indicateed only small variations. Our speech data consists of short utterances with a small number of syllables. The differences between several renditions of the same utterance produced clear differences for few speakers. Nevertheless, speech rate has been studied as a correlate of vocal age disguise (González Hautamäki *et al.*, 2017; Skoog Waller *et al.*, 2015; Waller and Eriksson, 2016) and is worth considering in the model. Specifically, speakers intending to sound older tend to decrease their speech rate, while speakers attempting to sound younger increase the speech rate.

The differences in formant frequencies and bandwidths were also inspected but visualizations did not produce consistent patterns that were associated with the voice conditions and speaker gender. Nevertheless, the variations between speakers is noticeable and were thus considered in the model of LLR scores. For example, Figure 4 depicts the formant variations, F1 and F2, for the vowel space of two speakers with low LLR scores. The ellipses represent the vowel space by voice condition. In comparison to the modal voice, the vowel spaces for disguised voices, are larger and shift downwards for the male speaker; for the female speaker, the vowel space for intended old voice is smaller and shift downwards; and the intended child's vowel space is larger and shift upwards. The formants variations are speaker -dependent and we can expect the vowel space to variate for every speaker too.

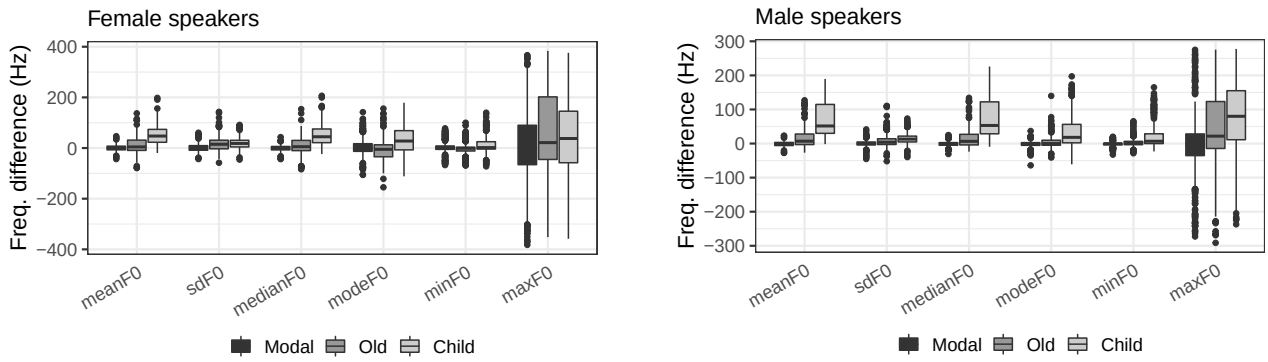
VIII. RESULTS: MIXED EFFECTS MODEL

We performed a statistical analysis to model the voice disguise effect on LLR scores. In Subsection VIII A, we analyze the calibrated LLR scores, our dependent variable, its distribution, and the effect caused by voice condition (modal, intended old, intended child). This analysis describes the effect of the voice condition and the variation introduced by the speaker effect. In Subsection VIII B, we explore how the change of the selected acoustic features explains the LLR score. The resulting model parameters are presented by the voice condition for comparative purposes, although voice condition is not included as fixed factor.

A. Effects of voice disguise on LLR scores

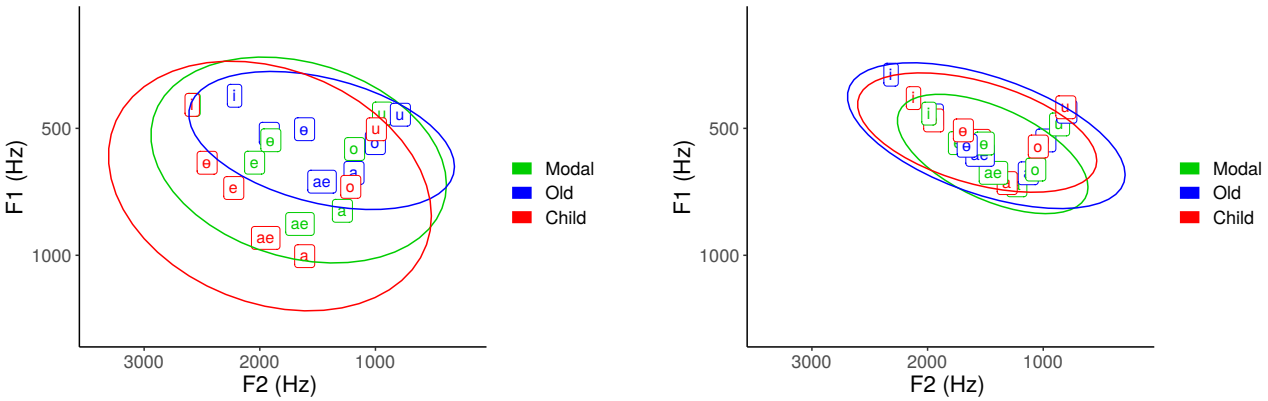
We first investigate whether the dependent variable (LLR score) is approximately normally distributed, which is an assumption in our model. This was verified by means of density plots and by quantile-quantile plots that should approximate a straight line. Through visual inspection, we concluded that even if the trials do not follow perfect normality, the assumption is reasonable. We therefore proceed to model the voice disguise effect in our LLR scores. We use the *lme4 package* (Bates *et al.*, 2015) to fit the linear mixed effects model.

For the next analysis, the voice condition is treated as an explanatory variable or *fixed effect* across the speakers and the spoken utterances. We can assume that the LLR scores vary for different voice conditions across speakers. The model can reflect these individual differences by assuming different random intercepts for each speaker. In addition to by-speaker variation, we expect a "random" variation between different sentences uttered by the same speaker. Speakers and utterances were then treated as random effects for the fixed effect of the voice condition. Table IV presents the parameters of the model, intercept and slopes for the linear model, for the voice condition effect on i-vector LLR scores. Results with x-vector LLR scores are included in the supplementary material[1].



(A) (B)

FIG. 3. Female (A) and male (B) speakers' F0 differences of summary statistics between trials' utterances presented by voice condition.



(A) (B)

FIG. 4. (Color online) An example of vowel space variation for a female (A) and male (B) speaker presented for modal, intended old and intended child voice conditions. The speakers were selected based on their low LLR scores for both i-vector PLDA and x-vector PLDA.

The *standard deviation* (SD) for random effects is a measure of how much variability in the dependent measure there is due to speakers and utterances. A similar variability by-speaker and by-sentence can be observed. Further, the variation between the disguised voices is similar for both conditions in the by-speaker analysis (for females in the i-vector system $SD = 2.84$ and 2.83). The *residual*, which corresponds to the variability that is neither due to speakers nor utterances, can be considered as an indication that each speakers uttered sentence has some factors that affect the ASV system score and are outside of this model. This is one motivation to include acoustical measures as fixed effects later on.

The coefficient for the intended old voice is the slope for that voice condition. For female speakers, for example, the coefficient −5.97 means that changing from modal to intended old voice causes the LLR score to go down by 5.97 units. In other words, the LLR score is lower in the intended old voice than for the modal voice and is even lower for the intended child voice (−8.39 relative to modal voice). So, from the speakers' point of view, the child voice is a more effective disguising strategy than the old voice role. Similar effects are observed for male speakers in both ASV systems.

We created a model without the effect of the voice condition on the ASV scores and compared it to the model that has that effect to analyze its importance. This can be performed using a standard likelihood test *ANOVA*. The *Akaike information criterion* (AIC) (Akaike, 1974) value was used to evaluate both models and identified the model with better fit. The AIC value decreases with better models. Considering the i-vector system's scores model results (See Table IV), we found that the voice condition affected the LLR score for

TABLE IV. Summary of the results for the mixed effects model of voice condition for calibrated i-vector ASV system's scores

		Female		Male	
Random effects:					
Groups	Voice	Variance	Std.Dev.	Variance	Std.Dev.
speaker	(Intercept)	0.97	0.98	4.25	2.06
	Old	8.09	2.84	13.36	3.65
	Child	8.03	2.83	35.94	5.99
sentence	(Intercept)	2.05	1.43	2.43	1.56
	Old	0.29	0.54	0.09	0.30
	Child	1.28	1.13	0.36	0.60
Residual		6.37	2.53	7.50	2.74

Fixed effects:						
	Estimate	Std. Error	t value	Estimate	Std. Error	t value
(Intercept)	4.42	0.46	9.77	7.03	0.60	11.82
Old	−5.97	0.56	−10.63	−8.04	0.71	−11.29
Child	−8.39	0.62	−13.44	−11.02	0.62	−9.64

female speakers ($\chi^2(12) = 1328$, $p < 2.2e^{-16}$), lowering it by about 5.97 ± 0.56 (standard errors) for the intended old voice and lowering about 8.39 ± 0.62 for the intended child voice. For male speakers, the variation by-speaker is higher than the by-sentence for the voice conditions, with a SD for the intended old voice of 3.65 and 5.99 for the intended child. The voice condition affected LLR score ($\chi^2(12) = 1623$, $p < 2.2e^{-16}$), lowering it by about 8.04 ± 0.71 (standard errors) for the intended old, and lowering even further 11.02 ± 0.62 for the intended child.

The voice condition effect on LLR scores is in line with the degradation in EER as displayed in Table II. In the subsequent analysis, we do not include the effect of voice condition, as a fixed effect, to explain the variation in LLR score. In this manner, we seek to understand how the acoustical features' differences can be associated to the trials' LLR scores without the explicit information of the voice condition.

B. Effects of acoustical variations on LLR scores

We performed a linear mixed effects analysis of the relationship between LLR scores, and the fixed effects: absolute differences of F0 summary statistics (mean, standard deviation, median, mode, minimum, and maximum), mean difference of formant frequencies, F1 to F4 and their bandwidths, B1 to B4, and difference of speech rate (syllables/second). Only the speakers were treated as random effects because the variance corresponding to sentences was small.

An important objective of this analysis is to identify the change of acoustical features that best explains the LLR score associated with the trial's enrollment and test utterances. We fitted a model by aggregating each feature difference at a time to select only those feature differences that were most significant from the 15 estimated ones. We searched for the highest Pearson correlation between the model fitted values and the LLR scores to compare between the models. After a feature difference was included in the model, we proceeded to evaluate the remaining feature differences, aggregating one-by-one and verifying the highest correlation to be included in the model. This process continued until there was no increase in the correlation and all the feature dif-

TABLE V. Coefficients (β) for the linear mixed effect model for female speakers presented by voice condition. Predictors for i-vector and x-vector same speaker scores: difference of mean F0, F3, and B4 between genuine trial utterances.

i-vector

	Modal	Old	Child	Pooled
β_0	5.90	0.97	−0.15	4.29
Feature				
dmeanF0 (β_1)	−0.11	−0.04	−0.05	−0.093
dmeanF3 (β_2)	−0.003	−0.005	−0.003	−0.007
dmeanB4 (β_3)	−0.02	−0.006	−0.002	−0.008

x-vector

β_0	9.79	1.67	1.02	7.86
Feature				
dmeanF0 (β_1)	−0.11	−0.04	−0.05	−0.11
dmeanF3 (β_2)	−0.003	−0.01	−0.005	−0.013
dmeanB4 (β_3)	−0.004	−0.004	−0.005	−0.012

TABLE VI. Coefficients (β) for the linear mixed effect model for male speakers presented by voice condition. Predictors for i-vector same speaker scores: difference of mean F0, F4, and B1. Predictors for x-vector same speaker scores: difference of median F0, F4, and F1.

i-vector

	Modal	Old	Child	Pooled
β_0	8.49	0.40	0.62	6.46
Feature				
dmeanF0 (β_1)	−0.11	−0.03	−0.04	−0.08
dmeanF4 (β_2)	−0.01	−0.002	−0.01	−0.01
dmeanB1 (β_3)	−0.005	−0.007	−0.01	−0.02

x-vector

β_0	10.07	0.36	−0.38	8.26
Feature				
dmedianF0 (β_1)	−0.13	−0.05	−0.10	−0.13
dmeanF4 (β_2)	−0.002	−0.01	−0.004	−0.02
dmeanF1 (β_3)	−0.001	−0.01	−0.002	−0.03

ferences had been included. The final model has the feature differences in a sequence that can be interpreted as a ranking in descending order that indicates the feature differences that contribute more information to the model. For simplicity, we chose the three top feature differences for the proposed model. The rest of the feature differences results are included in the supplementary material.

For the female model, the acoustical feature differences (fixed effects) that were more important for the fitted model were the differences of mean F0, mean F3, and mean B4. For the male model, the factors that were more explanatory for i-vector PLDA scores were the differences of mean F0, mean F4, and mean B1, while for x-vector PLDA were the differences of median F0, mean F4, and mean F1. Tables V and VI present the regression coefficients for the best models selection for female and male speakers respectively.

Table VII presents the Pearson correlations between the model fitted values and the LLR scores for both systems in the corresponding voice conditions. Further, Figures 5 and 6 depict the correlations between the model fitted values and the LLR scores for i-vector ASV system. The model based on the absolute difference of the selected features has a higher correlation for the disguised voices than for the modal voice. This indicates that the mixed effects model better explains the LLR score for the target trials with disguised voices, where we see more variation in the acoustical feature differences.

Table VIII presents the model parameters per gender and voice condition for the best fitted models. β_0 is the mean LLR for the trials in the respective gender and voice condition. σ^2 represents the variability that is not dependent on the speaker factor, while σ_b^2 describes the variability that is related to the speaker effect. The residual variance (σ^2) is high in all the cases, which indicates that even though the selected feature differences explain the LLR score per trial, there are factors that are not included in our model. The model can be potentially improved by including more acoustic features and/or other distance measures.

TABLE VII. Correlations for the model fitted values and the LLR scores (i-vector and x-vector ASV systems) corresponding to gender and voice conditions (modal, old and child). p-value $2.2e^{-16}$ with a confidence interval at 95%.

Model	i-vector	x-vector
Female – modal	0.51	0.57
Female – old	0.74	0.83
Female – child	0.74	0.80
Male – modal	0.64	0.67
Male – old	0.80	0.82
Male – child	0.87	0.92

TABLE VIII. Statistic parameters from the fitted models corresponding to gender and voice conditions (modal, old and child) for i-vector and x-vector ASV systems.

i-vector

Model	Population mean β_0	Residual errors variance σ^2	Variance speaker effects σ_b^2
Female – modal	5.90	6.14	0.77
Female – old	0.97	6.83	4.48
Female – child	−0.15	7.74	2.90
Male – modal	8.49	7.68	3.37
Male – old	0.40	8.95	2.99
Male – child	0.62	10.10	14.16

x-vector

Female – modal	9.79	8.09	1.98
Female – old	1.67	14.10	20.44
Female – child	1.02	13.72	13.81
Male – modal	10.07	8.08	4.70
Male – old	0.36	17.77	25.16
Male – child	−0.38	15.41	40.85

IX. DISCUSSION

We summarize our main observations as follows:

- **ASV performance degradation:** while our prior work (González Hautamäki *et al.*, 2017) on the AVOID corpus demonstrated severe perfor-

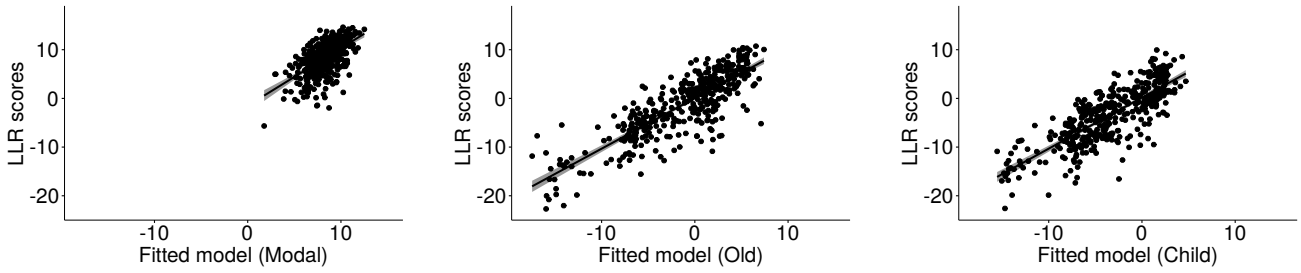


FIG. 5. Correlation for fitted model values and LLR scores (x-vector) for female speakers in their (A) modal ($r = 0.57$), (B) intended old ($r = 0.83$), and (C) intended child voices ($r = 0.80$).

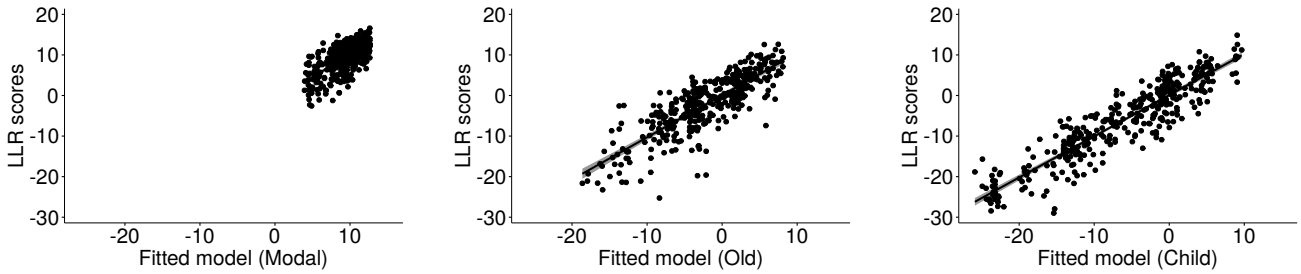


FIG. 6. Correlation for fitted model values and LLR scores (x-vector) for male speakers in their (A) modal ($r = 0.67$), (B) intended old ($r = 0.82$), and (C) intended child voices($r = 0.92$).

mance degradation of the i-vector system, we have confirmed that the x-vector approach is not immune to disguise either. The x-vector system systematically outperformed our i-vector system (Table II), but relative performance degradation for the former was worse. The mixed effects model indicates substantially lowered LLR scores from modal voice to the two disguised voices (Table IV).

- **Features with greatest explanatory power:** The feature ranking experiment (Tables V and VI) reveals the average (either mean or median) F0 difference to be the individually most important feature to explain the target LLR score. Importantly, this is consistent across the i-vector and the x-vector systems. F0 is followed by formant-related mismatches, whether in specific formant frequencies (F1, F3, F4) or bandwidths (B1, B4). Again, the observations are nearly consistent across the i-vector and the x-vector systems within each gender.

- **Mixed effect model as a whole:** How good is the mixed effect model at explaining the target LLR score as a whole? As Table VII and Fig. 6 indicate, the correlation between the fitted model with the selected explanatory features varies from 0.51 up to 0.92, with many values concentrated in $[0.7, 0.8]$. Despite the use of simple acoustic features, the correlations are deemed to be *strong* and indicate the usefulness of the proposed framework. Nonetheless, the large residual variances (Table VIII) indicate the presence of unmodeled effects, leaving scope for future improvements.

- **Modal vs. disguised voices:** The final model correlation (Table VII) is weaker for the modal than the two disguised voices (old, child). This is somewhat expected; for acoustically similar utterances (modal condition), ASV systems are tolerant against reasonable within-speaker variations by their design. However, when the trial utterances are acoustically very dissimilar (old and young conditions), the LLR degradation is stronger because the ASV systems are not designed to cope with these cases.

- **Model for i-vector vs. x-vector scores:** Table VII further indicates systematically higher correlations for the x-vector system. This is in line both with the larger relative degradation in EER (Table II) and the higher absolute values of the β

coefficients in Tables V and VI (on average). The x-vector system could be said to be more sensitive to large acoustic perturbations — though still outperforming the i-vector system.

The above findings suggest that the proposed interpretative framework is well-suited to cater for *acoustic explanations* to ASV performance degradation, especially under intentional and strong speech style changes. As our research data consists of speech produced by naive actors whose disguise strategies may be incompatible with each other, the authors are cautious to avoid over-interpreting observations relating to specific acoustic parameters. Nonetheless, one general observation was that a mismatch in average F0 and specific formants (frequencies and bandwidths) were the top features to explain target LLRs in both of the tested ASV systems. The commonality of the i-vector and the x-vector systems is their use of MFCC features that are extracted from the short-term power spectrum. The short-term power spectrum is notoriously sensitive to a number of distortions, including F0 mismatch (El-Jaroudi and Makhoul, 1991) and formant mismatch, which might explain the observed results.

X. CONCLUSIONS

We performed linear mixed effects analysis of the relationship between ASV system scores and acoustic and prosodic within-speaker differences. The considered fixed effects included voice condition (modal, intended old, intended child), difference of F0 statistics (mean, standard deviation, median, mode, minimum and maximum), mean difference of formants F1-F4, mean difference of their bandwidths B1-B4, and difference of speech rate (syllables/second). The random effect due to different speakers was considered as part of the variability of the model. Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity (homogeneity of variance) or normality. Feature selection for the final model was performed by aggregating the feature difference that added explanatory information to the model, which was defined by the highest correlation between the fitted values and the modeled LLR scores variable.

In principle, the ASV systems considered in this study use frame-level features — namely, MFCCs. Nonetheless, we did not observe a drastic effect caused by differences in formants, but it is the opposite case for F0. In our previous work (González Hautamäki *et al.*, 2017), we already noted the drastic F0 modifications implemented by the speakers when they attempted to sound like an elderly person and a child. The new insight derived from the present work is the explicit modeling of the link between acoustic changes and degraded target LLR scores. While our findings are specific to the selected dataset and recognizer, the same mixed effects model approach could be used to analyze the potential sensitivity of other ASV systems to change in other features as well. We envision that modeling speaker dependency in, for example, trial-based calibration, can significantly improve the calibration performance because all trials are assumed to be independent in current systems.

The results suggest potential future improvements to state-of-the-art ASV. Perhaps x-vector systems could be made more robust against within-speaker speech style variation by replacing MFCCs (or the power spectrum estimator) with alternative methods designed to tackle specifically identified types of acoustic mismatch (such as F0 difference). Another path could be novel data augmentation strategies to enlarge target speaker's enrollment data with acoustically-manipulated versions.

ACKNOWLEDGMENTS

This research was partially funded by the Academy of Finland, projects no. 309629 and 313970. We are thankful to Ville Vestman for his assistance with the x-vector PLDA system.

[1]See Supplementary materials at [URL will be inserted by AIP] for additional results and analysis for the proposed linear mixed effects model and acoustical variations of formant frequencies for selected speakers.

Adami, A. (**2007**). "Modeling prosodic differences for speaker recognition," Speech Communication **49**(4), 277–291.

Ajili, M. (**2017**). "Reliability of voice comparison for forensic applications. (fiabilité de la comparaison des voix dans le cadre judiciaire)," Ph.D. thesis, University of Avignon, France.

Ajili, M., Bonastre, J.-F., and Rossato, S. (**2018**). "Voice Comparison and Rhythm: Behavioral Differences between Target and Non-target Comparisons," in *Interspeech 2018*, ISCA, pp. 1061–1065, http://www.isca-speech.org/archive/Interspeech_2018/abstracts/0061.html, doi: 10.21437/Interspeech.2018-61.

Akaike, H. (**1974**). "A new look at the statistical model identification," IEEE Transactions on Automatic Control **19**(6), 716–723.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," Journal of Statistical Software **67**(1), 1–48.

Boersma, P. (**1993**). "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proc. of the Institute of Phonetic Sciences*, Vol. 17, pp. 97–110.

Boersma, P., and Weenink, D. (**2015**). "Praat: doing phonetics by computer [Computer program]" Version 5.4.09, retrieved 15 June 2015 from http://www.praat.org/.

Brümmer, N., Burget, L., Cernocky, J., Glembek, O., Grezl, F., Karafiat, M., van Leeuwen, D. A., Matejka, P., Schwarz, P., and Strasheim, A. (**2007**). "Fusion of heterogeneous speaker recognition systems in the stbu submission for the nist speaker recognition evaluation 2006," IEEE Transactions on Audio, Speech, and Language Processing **15**(7), 2072–2084.

Casella, G., and Berger, R. L. (**2002**). *Statistical inference*, **2** (Duxbury Pacific Grove, CA).

Childers, D. (**1978**). IEEE Press selected reprint series *Modern Spectrum Analysis* (New York, IEEE Pr.), pp. 34–41.

Chung, J. S., Nagrani, A., and Zisserman, A. (**2018**). "Voxceleb2: Deep speaker recognition," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018.*, pp. 1086–1090.

Cohen, J. (**1988**). *Statistical power analysis for the behavioral sciences*, **2** (Lawrence Earlbaum Associates, USA).

De Jong, N. H., and Wempe, T. (**2009**). "Praat script to detect syllable nuclei and measure speech rate automatically," Behavior Research Methods **41**(2), 385–390.

Dehak, N., Dumouchel, P., and Kenny, P. (**2007**). "Modeling prosodic features with joint factor analysis for speaker verification," IEEE Transactions on Audio, Speech, and Language Processing **15**(7), 2095–2103, doi: 10.1109/TASL.2007.902758.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., and Ouellet, P. (**2011**). "Front-end factor analysis for speaker verification," IEEE Transactions on Audio, Speech, and Language Processing **19**(4), 788–798.

Dellwo, V., Leemann, A., and Kolly, M.-J. (**2012**). "Speaker idiosyncratic rhythmic features in the speech signal," in *Interspeech*, pp. 1584–1587.

Dellwo, V., Leemann, A., and Kolly, M.-J. (**2015**). "Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors," The Journal of the Acoustical Society of America **137**(3), 1513–1528, http://asa.scitation.org/doi/10.1121/1.4906837, doi: 10.1121/1.4906837.

Dempster, A., Laird, N., and Rubin, D. (**1977**). "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society. Series B **39**(1), 1–38.

Doddington, G. R., Przybocki, M. A., Martin, A. F., and Reynolds, D. A. (**2000**). "The NIST speaker recognition evaluation–overview, methodology, systems, results, perspective," Speech Communication **31**(2), 225–254.

El-Jaroudi, A., and Makhoul, J. (**1991**). "Discrete all-pole model," IEEE Transactions on Signal Processing **39**(2), 411–423.

Farrús, M., Hernando, J., and Ejarque, P. (**2007**). "Jitter and shimmer measurements for speaker recognition," in *Proc. Interspeech*, Belgium, pp. 778 – 781.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., and Zue, V. (**1993**). "TIMIT acoustic-phonetic continuous speech corpus LDC93S1" Web Download, linguistic Data Consortium, Philadelphia.

González Hautamäki, R. (**2017**). "Human-induced voice modification and speaker recognition: automatic, perceptual and acoustic perspectives," Ph.D. dissertation, University of Eastern Finland. Dissertations in Forestry and Natural Sciences, 290, Joensuu, Finland, table 6.5, pag. 56.

González Hautamäki, R., Kanervisto, A., Hautamäki, V., and Kinnunen, T. (**2018**). "Perceptual evaluation of the effectiveness of voice disguise by age modification," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, pp. 320–326, http://dx.doi.org/10.21437/Odyssey.2018-45, doi: 10.21437/Odyssey.2018-45.

González Hautamäki, R., Sahidullah, M., Hautamäki, V., Bentz, M., Werner, S., and Kinnunen, T. (**2018**). "Corpus of age-related voice disguise (AVOID)" http://urn.fi/urn:nbn:fi:lb-2018060621, available on Kielipankki – The Language Bank of Finland.

González Hautamäki, R., Sahidullah, M., Hautamäki, V., and Kinnunen, T. (**2017**). "Acoustical and perceptual study of voice disguise by age modification in speaker verification," Speech Communication **95**, 1–15.

González Hautamäki, R., Sahidullah, M., Kinnunen, T., and Hautamäki, V. (**2016**). "Age–related voice disguise and its impact in speaker verification accuracy," in *Proc. Odyssey: the Speaker and Language Recognition Workshop*, pp. 277–282.

Greenberg, C. S., Martin, A. F., Barr, B. N., and Doddington, G. R. (**2011**). "Report on performance results in the NIST 2010 speaker recognition evaluation," in *Proc. Interspeech*, Florence, Italy, pp. 261–264.

Hanilci, C., Kinnunen, T., Saeidi, R., Pohjalainen, J., Alku, P., and Ertas, F. (**2013**). "Speaker identification from shouted speech: Analysis and compensation," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8027–8031.

Hansen, J. H., and Hasan, T. (**2015**). "Speaker recognition by machines and humans: a tutorial review," IEEE Signal Processing Magazine **32**(6), 74–99.

Hansen, J. H. L., Nandwana, M. K., and Shokouhi, N. (**2017**). "Analysis of human scream and its impact on text-independent speaker verification," The Journal of the Acoustical Society of America **141**(4), 2957–2967, http://asa.scitation.org/doi/10.1121/1.4979337, doi: 10.1121/1.4979337.

Hatch, A. O., Kajarekar, S. S., and Stolcke, A. (**2006**). "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. Interspeech*, ISCA, pp. 1471 – 1474.

Kahn, J., Audibert, N., Rossato, S., and Bonastre, J. (**2010**). "Intra-speaker variability effects on speaker verification performance," in *Odyssey 2010: The Speaker and Language Recognition Workshop, Brno, Czech Republic, June 28 - July 1, 2010*, p. 21.

Larcher, L., Lee, K., Ma, B., and Li, H. (**2014**). "Text-dependent speaker verification: Classifiers, databases and RSR2015," Speech Communication **60**, 56–77.

Lee, K. A., Larcher, A., Wang, W., Kenny, P., Brummer, N., van Leeuwen, D. A., Aronowitz, H., Kockmann, M., Vaquero, C., Ma, B., Li, H., Stafylakis, T., Alam, J., Swart, A., and Perez, J. (**2015**). "The RedDots data collection for speaker recognition," in *"Proc. Interspeech"*, Dresden, Germany, pp. 2996–3000.

Leemann, A., and Kolly, M.-J. (**2015**). "Speaker-invariant suprasegmental temporal features in normal and disguised speech," Speech Communication **75**, 97–122.

Leemann, A., Kolly, M.-J., and Dellwo, V. (**2014**). "Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison," Forensic Science International **238**, 59–67.

Lei, Y., and Hansen, J. H. (**2016**). "Corpora for the evaluation of robust speaker recognition systems," in *"Proc. Interspeech"*, San Francisco, USA, pp. 2776–2780.

Mandasari, M. I., Saeidi, R., and van Leeuwen, D. A. (**2015**). "Quality measures based calibration with duration and noise dependency for speaker recognition," Speech Communication **72**, 126–137.

Mary, L., and Yegnanarayana, B. (**2008**). "Extraction and representation of prosodic features for language and speaker recognition," Speech Communication **50**(10), 782–796.

Moez, A., Jean-Francois, B., Waad, B. K., Solange, R., and Juliette, K. (**2016**). "Phonetic content impact on Forensic Voice Comparison," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, San Diego, CA, pp. 210–217, http://ieeexplore.ieee.org/document/7846267/, doi: 10.1109/SLT.2016.7846267.

Park, S. J., Yeung, G., Vesselinova, N., Kreiman, J., Keating, P. A., and Alwan, A. (**2018**). "Towards understanding speaker discrimination abilities in humans and machines for text-independent short utterances of different speech styles," The Journal of the Acoustical Society of America **144**(1), 375–386, http://asa.scitation.org/doi/10.1121/1.5045323, doi: 10.1121/1.5045323.

Patterson, H. D., and Thompson, R. (**1971**). "Recovery of inter-block information when block sizes are unequal," Biometrika **58**(3), 545–554.

Pietrowicz, M., Hasegawa-Johnson, M., and Karahalios, K. G. (**2017**). "Acoustic correlates for perceived effort levels in male and female acted voices," The Journal of the Acoustical Society of America **142**(2), 792–811, http://asa.scitation.org/doi/10.1121/1.4997189, doi: 10.1121/1.4997189.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (**2011**). "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, IEEE Signal Processing Society, iEEE Catalog No.: CFP11SRW-USB.

Prince, S. J. D., and Elder, J. H. (**2007**). "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. of International Conference on Computer Vision (ICCV)*, IEEE, pp. 1–8.

Rodman, R., and Powell, M. (**2000**). "Computer recognition of speakers who disguise their voice," in *Proc. of the Int. Conf. on Signal Processing Applications and Technology ICSPAT*.

Saeidi, R., Huhtakallio, I., and Alku, P. (**2016**). "Analysis of face mask effect on speaker recognition," in *"Proc. Interspeech"*, pp. 1800–1804.

Schmidt-Nielsen, A., and Stern, K. R. (**1985**). "Identification of known voices as a function of familiarity and narrowband coding," The Journal of the Acoustical Society of America **77**(2), 658–663.

Shriberg, E., Ferrera, L., Kajarekar, S., Venkataraman, A., and Stolcke, A. (**2005**). "Modeling prosodic feature sequences for speaker recognition," Speech Communication **46**(3–4), 455–472.

Skoog Waller, S., Eriksson, M., and Sörqvist, P. (**2015**). "Can you hear my age? influences of speech rate and speech spontaneity on estimation of speaker age," Frontiers in Psychology **6**(978).

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (**2018**). "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE ICASSP*, Calcary, Canada, pp. 5329–5333.

Sönmez, M. K., Heck, L. P., Weintraub, M., and Shriberg, E. (**1997**). "A lognormal tied mixture model of pitch for prosody based speaker recognition," in *Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997, Rhodes, Greece, September 22-25, 1997*.

Vestman, V., Gowda, D., Sahidullah, M., Alku, P., and Kinnunen, T. (**2018**). "Speaker recognition from whispered speech: A tutorial survey and an application of time-varying linear prediction," Speech Communication **99**, 62–79, https://linkinghub.elsevier.com/retrieve/pii/S0167639317302637, doi: 10.1016/j.specom.2018.02.009.

Waller, S., and Eriksson, M. (**2016**). "Vocal age disguise: The role of fundamental frequency and speech rate and its perceived effects," Frontiers in Psychology **7**(1814).

Wang, D., and Narayanan, S. S. (**2007**). "Robust speech rate estimation for spontaneous speech," IEEE Transactions on Audio, Speech, and Language Processing **15**(8), 2190–2201.

Zhang, C. (**2012**). "Acoustic analysis of disguised voices with raised and lowered pitch," in *Proc. Int. Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 353–357.

Zhang, C., and Tan, T. (**2008**). "Voice disguise and automatic speaker recognition," Forensic Science International **175**(23), 118–122.