



HAL
open science

Shouted and whispered speech compensation for speaker verification systems

Santi Prieto, Alfonso Ortega, Iván López-Espejo, Eduardo Lleida

► To cite this version:

Santi Prieto, Alfonso Ortega, Iván López-Espejo, Eduardo Lleida. Shouted and whispered speech compensation for speaker verification systems. *Digital Signal Processing*, 2022, 127, pp.103536. 10.1016/j.dsp.2022.103536 . hal-04886779

HAL Id: hal-04886779

<https://hal.science/hal-04886779v1>

Submitted on 14 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Shouted and Whispered Speech Compensation for Speaker Verification Systems

Santi Prieto^a, Alfonso Ortega^b, Iván López-Espejo^c, Eduardo Lleida^b

^a*VeriDas | das-Nano, Navarre, Spain*

^b*ViVoLab, Aragón Institute for Engineering Research (I3A)
University of Zaragoza, Spain*

^c*Department of Electronic Systems, Aalborg University, Denmark*

Abstract

Nowadays, speaker verification systems begin to perform very well under normal speech conditions due to the plethora of neutrally-phonated speech data available, which are used to train such systems. Nevertheless, the use of vocal effort modes other than normal severely degrades performance because of vocal effort mismatch. In this paper, in which we consider whispered, normal and shouted speech production modes, we first study how vocal effort mismatch negatively affects speaker verification performance. Then, in order to mitigate this issue, we describe a series of techniques for score calibration and speaker embedding compensation relying on logistic regression-based vocal effort mode detection. To test the validity of all of these methodologies, speaker verification experiments using a modern x-vector-based speaker verification system are carried out. Experimental results show that we can achieve, when combining score calibration and embedding compensation relying upon vocal effort mode detection, up to 19% and 52% equal error rate (EER) relative improvements under the shouted-normal and whispered-normal scenarios, respectively, in comparison with a system applying neither calibration nor compensation. Compared to our previous work [1], we obtain a 7.3% relative improvement in terms of EER when adding score calibration in shouted-normal All *vs.* All condition.

*

Email addresses: sprieto@veridas.com (Santi Prieto), ortega@unizar.es (Alfonso Ortega), ivl@es.aau.dk (Iván López-Espejo), lleida@unizar.es (Eduardo Lleida)

Keywords: Speaker verification, vocal effort mismatch, shouted speech, whispered speech, domain compensation, embedding compensation, deep learning.

1. Introduction

Speaker verification systems are becoming very popular and highly used in everyday life applications due to their improved performance and the reduced amount of speech required to enroll/verify users [2, 3, 4]. In part, this is due to the increased availability of high computational power and large amounts of speech data that have made the use of deep learning techniques possible [5, 6, 7, 8]. However, despite the fact that the accuracy of this technology is considerably high, there are many scenarios in which deep models can present problems such as when used in situations where speech is not neutrally phonated.

Human speech can have different modes of production depending on physical and emotional conditions. The vocal apparatus can be seen as a musical instrument that every person plays when she/he is speaking with many tinges. For this reason, speaker verification systems do not work as well when facing fragments with different vocal effort modes than when facing two neutrally-phonated speech fragments. Fundamentally, five different vocal effort modes are considered: whispered, soft, neutral or normal, loud and shouted speech with different characteristics that can adversely affect the performance of speaker verification systems [9].

A little number of works has studied how vocal effort affects automatic speech recognition, e.g., [10, 11]. For example, the authors of [11] conclude that when whispered speech is considered, a 50% decrease in performance can be expected. Besides, in [12], a study on shouted *versus* normal speech in the context of speaker verification is carried out. While the accuracy of the system with normal speech is 94.7%, that decreases to 44.7% when shouted speech is employed. Similarly, in [13], a drop in performance from 100% to only 8.7% in speaker identification is reported, so a shouted speech detection and compensation method

is proposed. For the same task, the authors of [14] apply both mixture probabilistic linear discriminant analysis (mix-PLDA) and trial-based calibration with condition PLDA similarity (TBC-CPLDA) to improve system robustness with 18% and 33% relative improvement in discrimination and calibration performance, respectively, on the SRI-FRTIV (Five-way Recorded Toastmaster Intrinsic Variation) corpus. In this same vein, [15] demonstrated very recently that speech produced under the Lombard effect can degrade speaker verification performance significantly in comparison with neutrally-phonated speech typically present in “clean” acoustic conditions. It was found that the extent of the degradation depends on the noise type, with large crowd noise having the biggest impact —i.e., 7.5% absolute increase of equal error rate (EER)—. A quality measure function-based calibration approach, incorporating the noise level as an additional term in linear score calibration, was found to outperform conventional calibration across all noise types. Besides, Sarria-Paja and Falk [5] proposed in 2017 three innovative types of speech features to take into account complementary characteristics of whispered and normal speech. By training three different speaker verification systems with their proposed speech features, they obtained EER improvements of 66% and 63% for normal and whispered speech, respectively, over a baseline using traditional Mel-frequency cepstral coefficients (MFCCs). Then, in [16], Vestman *et al.* studied whispered speech, introducing a new modeling technique that involves a long-term speech analysis based on a joint utilization of frequency domain linear prediction and time-varying linear prediction (FDLP-TVLP). For the experiments, they used the CHAINS (CHARacterizing INdividual Speakers) corpus [17] to test whispered-normal mismatch conditions and demonstrated that FDLP-TVLP features improve speaker recognition EER by 7–10% over standard MFCC features. Another method to solve the vocal effort mismatch between normal and whispered speech in speaker verification [6] proposes fusing modulation spectral features with bottleneck features at a feature- and a score-level, obtaining a 68-70% improvement with respect to an i-vector-based baseline system. Another approach [7] proposes a feature mapping between whispered MFCC features and normal

MFCC features by maximizing cosine similarity between neutral and whisper i-vectors, getting a 24% relative improvement. Finally, formant and formant-gap features achieved, over other speech features like MFCCs, an absolute 3.79% EER improvement when dealing with whispered-normal mismatch conditions in the context of an i-vector/PLDA speaker verification system [8].

Closely related to the vocal effort question, the emotional state of a speaker also conditions the way speech is produced. Therefore, intra-speaker variability as a result of different emotions (e.g., happiness and anger) can challenge speaker verification systems [18] in a similar way as vocal effort mismatch does. For example, the authors of [19] show how the performance of an i-vector-based speaker verification system trained with normal speech data is negatively affected by speech produced under different emotions. While existing emotion-invariant speaker verification approaches may be applied to deal with the vocal effort mismatch problem raised in this work, they either are not suitable for modern neural network-based speaker verification [20] or have some drawbacks [21, 22]. With respect to these drawbacks, [22] involves training speaker-dependent Gaussian mixture models (GMMs), whereas we think that a speaker-independent methodology is desirable. Furthermore, [21] proposes an emotion-invariant speaker embedding extractor that requires enough emotional speech data to be trained. However, vocal effort speech data scarcity prevents us from considering such an approach, so other smart techniques are devised in the present paper.

In this work, we study whispered, normal and shouted modes in order to better characterize them and show why deep models do not work as well when speaker verification systems face phonation modes other than normal speech. Before carrying out any compensation, we present a simple shouted-normal and whispered-normal detector to classify speech segments into those three different categories. After that, we introduce a score calibration method that improves the performance when applied to the verification of speech segments produced in different vocal effort conditions. Finally, we propose compensation techniques, in the speaker embedding (i.e., x-vector in this work) domain, to reduce the vocal

effort mismatch between embeddings, regardless of the mode of production.
90 Compared to our previous work [1], we obtain a 7.3% relative improvement
in terms of EER when adding score calibration in shouted-normal All *vs.* All
condition.

The remainder of this paper is organized as follows: a brief study of whis-
pered, normal and shouted speech is presented in Section 2. In Section 3, the
95 experimental framework is explained. Then, a simple shouted and whispered
speech detection method is described in Section 4. Section 5 illustrates the score
calibration method, while Section 6 is dedicated to the compensation techniques
proposed in this work. Experimental results are presented in Section 7. Finally,
Section 8 concludes this work.

100 **2. On Vocal Effort Modes: Normal, Whispered and Shouted**

In this section, we briefly review some of the most prominent characteristics
of the different vocal effort modes considered in our study: normal, whispered
and shouted. This review will be illustrated by Figure 1, which shows the log-
magnitude spectra of a same phrase (in Spanish) uttered by a male speaker in
105 normal, whispered and shouted modes.

Among all the possible vocal effort modes, whispered speech production is
the lowest one in terms of energy [9], and it is mainly characterized by an
aperiodic weak excitation [13] due to almost non-existing vocal cord vibration
[23]. Thus, this is, in turn, the vocal mode with the lowest sound energy level [9].
110 While the normal speech production mode can be considered to lie somewhere
in the middle, at the other end of the range we can find the shouted vocal effort
mode. The latter is characterized by both a fast variation of the period of the
vocal cords and a notable voice excitation [13]. As a result, the shouted vocal
mode is the one with the highest sound energy level [9]. Such a sound energy
115 level increase along the whispered-normal-shouted speech production modes can
be observed at a glance in Figure 1.

Other acoustic features such as, e.g., spectral tilt, pitch and formant fre-

frequencies are also affected by the speech production mode due to its particular
 vocal tract resonance configuration [24]. In [9], Zhang and Hansen conclude
 120 that the spectral tilt decreases for both whispered and shouted modes with
 respect to normal vocal effort. While this might be attributed to a relatively
 greater contribution of the consonants to the total whispered speech energy, in
 the shouted speech case this could be as a result of the glottal pulse with more
 regular shapes [9].

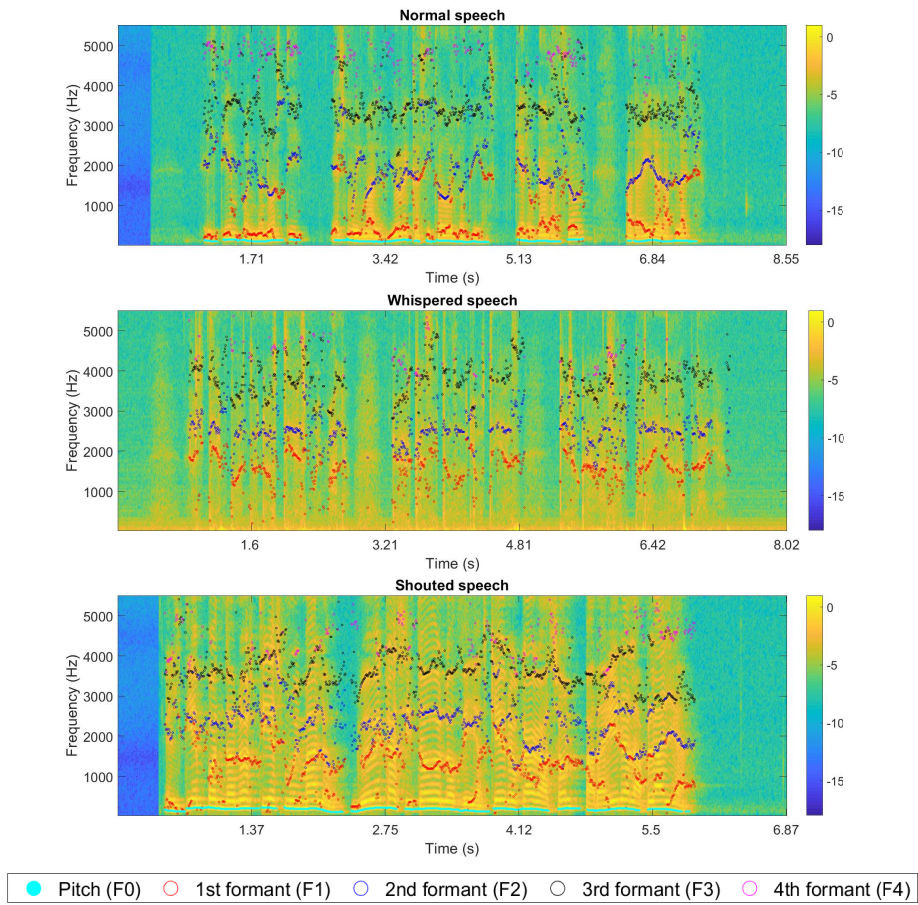


Figure 1: Log-magnitude spectra of a same phrase (in Spanish) uttered (by a male speaker),
 from top to bottom, in normal, whispered and shouted modes. Estimates of pitch and the
 first four formants obtained by Praat [25] are overlaid on these spectra.

125 Considering the fundamental frequency (F0) variation, it is well-known that

Pitch/Formant	Normal	Whispered	Shouted
Pitch (F0)	121 ± 13	—	187 ± 15
1st formant (F1)	477 ± 210	1,608 ± 242	1,275 ± 336
2nd formant (F2)	2,002 ± 405	2,526 ± 181	2,375 ± 308
3rd formant (F3)	3,449 ± 234	3,804 ± 276	3,595 ± 306
4th formant (F4)	4,800 ± 175	4,388 ± 292	4,552 ± 331

Table 1: Vocal mode-dependent fundamental frequency (F0) and formant frequency values (from F1 to F4), in Hz, obtained as the median of the time sequences of fundamental frequency and formant frequency estimates displayed in Figure 1. Fundamental frequency and formant frequency values are presented along with their corresponding median absolute deviation values.

it tends to increase with higher vocal effort [26, 27] because of the increased subglottal pressure [28] and other factors [24]. Fundamental frequency estimates obtained by using Praat [25] are overlaid on the spectral representations in Figure 1. For the sake of comparison, Table 1 shows vocal mode-dependent F0 values, in Hz, obtained as the median of the time sequences of F0 estimates displayed in Figure 1. Fundamental frequency values in Table 1 are presented along with their corresponding median absolute deviation values. At this point, it should be noted that we use the sample median instead of the sample mean due to the further robustness of the former for the estimation of the central value of a probability distribution. Whereas the F0 estimate for whispered speech is not provided due to the absence of vocal folds vibration in this vocal mode, we can see, according to Table 1, that F0 is clearly greater for shouted than for normal speech, as expected.

Regarding formant structure, it was found that (mainly) the first and second formant frequencies (F1 and F2) and all formant bandwidths tend to increase in whispered with respect to normal speech [29, 30]. Similarly, in [27], Elliott states that the first formant frequency (F1) increases in shouted speech compared to normal speech while the second and third formant frequencies (F2 and F3) do not change much. As for the F0 case, formant frequency estimates also obtained

145 using Praat are overlaid on the spectrograms in Figure 1. Again, for the sake
of comparison, Table 1 also shows vocal mode-dependent formant frequency
values —along with their corresponding median absolute deviation values—,
in Hz, estimated as the median of the time sequences of formant frequency
estimates depicted in Figure 1. As it can be seen, the aforementioned formant
150 structure alterations across vocal effort modes are very much endorsed by the
illustrative example of Figure 1.

Let us also observe that different emotions provoke changes in terms of speech
production. For example, and as it comes to no surprise, speech produced under
stress conditions (e.g., in life-threatening scenarios) and angry speech show some
155 similarities with shouted speech. Particularly, all these types of speech exhibit
increased F0 and intensity with respect to neutral speech [31, 32, 33] due to
higher vocal effort. However, while angry speech also involves an increase of
the first formant frequency as for shouted speech, the latter does not show a
relevant increase of the second formant frequency as angry speech does [34].

160 To sum up, whispered, normal and shouted speech can be primarily differ-
entiated in terms of sound energy level, which is directly correlated with speech
fundamental frequency [26, 27]. Furthermore, whispered and shouted speech
exhibits increased first formant frequencies with respect to normal speech, and
whispered speech shows a higher second formant frequency than the other two
165 modes. Since speaker representations for speaker verification are derived from
the speech signal, they are also impacted by the vocal effort mode, which might
lead to poor speaker verification performance in case of vocal effort mode mis-
match. This question is discussed in the remainder of the paper.

3. Experimental Setup

170 In this section, we describe the experimental setup. First, in Subsection
3.1, we explain the speaker verification system used to obtain the baseline ex-
periment in this work. Second, in Subsection 3.2, we analyze the effects of
vocal effort in the embedding domain. Then, we divide the experiments into

normal-shouted and normal-whispered speech due to the different characteristics of the datasets that are described in Subsection 3.3. Finally, we explain the whole experimental framework in Subsection 3.4, which includes the different modules that are considered to improve the performance of the baseline speaker verification system when vocal effort mismatch is present.

3.1. Baseline Speaker Verification System

In this subsection, we explain the characteristics of the speaker verification system used to perform the baseline experiments carried out with the datasets described in Subsection 3.3. In Figure 2, we can see the block diagram of this speaker verification system.

The feature extractor module receives the input signal and returns 30-dimensional MFCC [35] vectors. In particular, the speech signal is framed employing a 25 ms analysis window with a 10 ms shift. Furthermore, this module uses a voice activity detector to discard non-speech frames.

The time-delay neural network (TDNN) module is fed with MFCC features computed from a single-speaker utterance to output a 512-dimensional speaker embedding (x-vector) that represents the speaker present in the input utterance. The reason behind using a TDNN [36] for speaker embedding extraction lies in the popularity of this architecture for the speaker verification task [37, 38]. Success of TDNNs comes from their ability to effectively modeling long-range dependencies in time sequences like speech signals [36, 37, 38].

In Table 2 [37], we describe the TDNN configuration for an input segment with T frames. The first five layers work on speech frames with a temporal context centered at frame t . The statistics pooling layer aggregates all the T frame-level outputs from the fifth layer and computes its mean and standard deviation. The statistics are 1,500-dimensional vectors, calculated once for each input segment. This process aggregates information across the time dimension in such a manner that the subsequent layers work on the whole segment (layer context of $\{0\}$ and total context of T). The mean and standard deviation are concatenated and passed through segment-level layers followed by an output

Layer	Layer context	Total context	Input \times output
frame1	$[t-2, t+2]$	5	120×512
frame2	$\{t-2, t, t+2\}$	9	$1,536 \times 512$
frame3	$\{t-3, t, t+3\}$	15	$1,536 \times 512$
frame4	$\{t\}$	15	512×512
frame5	$\{t\}$	15	$512 \times 1,500$
stats pooling	$[0, T)$	T	$1,500T \times 3,000$
segment6	$\{0\}$	T	$3,000 \times 512$
segment7	$\{0\}$	T	512×512
softmax	$\{0\}$	T	$512 \times N$

Table 2: Architecture of the time-delay neural network (TDNN) used in this work for speaker embedding (x-vector) extraction [37].

softmax layer. All the non-linearities are rectified linear units (ReLUs) [37].

205 Once the TDNN has been trained, x-vectors are extracted from the affine component of layer “segment6”. At runtime, this architecture has a total of 4.2 million parameters [37].

Finally, the PLDA back-end module is used to obtain scores for each verification trial by comparing two different embeddings. Before PLDA scoring, 210 x-vectors are centered, whitened, reduced in terms of dimensionality by means of linear discriminant analysis (LDA) and length-normalized. Let us notice that, after application of LDA, the number of components of the vectors employed by PLDA is 150.

In our experimental setup, the primary metric to evaluate the performance 215 of the systems is equal error rate (EER). In a detection system, EER is the value in which false acceptance rate (FAR) and false rejection rate (FRR) are equal. FAR is the proportion of unauthorized users misclassified by the system, and FRR is the fraction of genuine users misclassified as unauthorized users by the system.

220 Our baseline speaker verification system is implemented according to the

x-vector-based Kaldi [39] recipe using augmented versions of the VoxCeleb1 [40] and VoxCeleb2 [41] datasets¹, and, for the sake of reproducibility, freely available models are downloaded². The EER obtained using this baseline system is 3.1% [39].

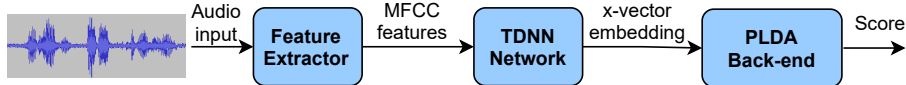


Figure 2: Block diagram of the speaker verification system used for experimental purposes in this paper.

225 3.2. Vocal Effort Mode Impact on X-Vectors

The characteristics of vocal effort modes described in Section 2 also affect the embeddings extracted using the models trained with normal speech. To show the effects in the embedding domain, we use t-SNE (*t*-distributed Stochastic Neighbor Embedding) [42], which allows us to visualize, in a two-dimensional
 230 space, 512-component embeddings. We can see in Figure 3a four different embedding clusters that tend to be grouped by gender and vocal effort condition, in this case, normal and shouted speech. In the speaker verification task, the same speaker with different vocal effort condition will not be correctly verified due to this mismatch not appropriately compensated by the embedding extractor.

In Figure 3b, the same behavior shown in Figure 3a can be appreciated but
 235 considering whispered speech this time. There are four different clusters that again tend to be grouped by gender and vocal effort condition corresponding to whispered and normal speech, male and female speakers. Similarly, this effect degrades the performance of the speaker verification system, i.e., when different
 240 vocal effort conditions are involved. In summary, this fact justifies the need for techniques able to mitigate the adverse effects of vocal effort mismatch on the performance of speaker verification systems.

¹<https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb>

²<https://kaldi-asr.org/models/m7>

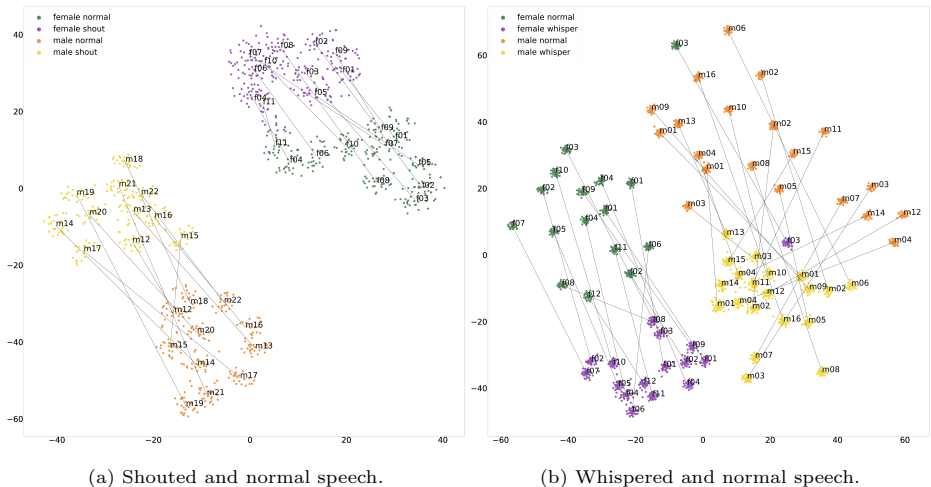


Figure 3: Speaker embedding representations obtained by means of t-SNE [42].

3.3. Shouted and Whispered Speech Test Datasets

In this subsection, we explain the databases used for experimental purposes.

245 In this work, we have experimented with two databases: normal-shouted and normal-whispered. Notice that, apart from the fact that these two databases are independent, the main reason we do not consider the shouted-whispered condition is because the goal of applying embedding compensation (from shouted to normal and whispered to normal) is to improve the performance of speaker

250 verification systems trained using normal speech only. Besides, due to data scarcity, shouted and whispered speech detection, score calibration and embedding compensation experiments are carried out using leave-one-speaker-out cross-validation. All the utterances in the corpora are processed to extract x-vectors according to what outlined in Subsection 3.1 and further detailed in

255 [37].

The shouted speech corpus used to perform the shouted *versus* normal speech experiments is the one presented in [13]. It consists of 11 male and 11 female speakers. Each of them recorded 24 sentences speaking with neutral phonation and the same 24 sentences shouting. The sentences were recorded in an anechoic

260 chamber using a high-quality microphone. Channel effects and environment

variations were completely excluded. The sentences were spoken in Finnish, half in imperative and half in indicative mode. The average duration of each utterance is 3 seconds. Four different conditions are considered for experimental evaluation:

- 265 • **All vs. All (A-A)**: All the shouted and normal speech utterances are verified against all the rest, which yields a total of 557,040 verification trials.
- **Normal vs. Normal (N-N)**: Normal speech utterances are verified against the rest of neutrally-phonated utterances, which yields a total of
270 139,128 verification trials.
- **Shouted vs. Shouted (S-S)**: Shouted speech utterances are verified against the rest of shouted utterances, which yields 139,128 verification trials.
- **Normal vs. Shouted (N-S)**: Normal speech utterances are verified
275 against shouted speech utterances, which yields a total of 278,784 verification trials.

For whispered speech, the CHAINS corpus [17], specifically developed to facilitate the task of speaker recognition, was used. This dataset is composed of recordings from 36 speakers, using distinct and well-defined speech registers:
280 neutral speech, synchronized with a co-speaker, synchronized and repetitively, whispered and at a rapid pace. Speakers use 17 different dialects of English, and 28 of the announcers (14 women and 14 men) come from the East of Ireland. The remaining 8 (4 women and 4 men) are from the United States or the United Kingdom. For each speaker, we have 37 whispered speech recordings plus other
285 37 neutral speech recordings. Four of the 37 recordings are fragments of fables and, therefore, have a long duration, i.e., between 30 seconds and 1 minute. The rest of them are short sentences, i.e., between 2 and 5 seconds. For our experiments, audios between 30 seconds and 1 minute have been removed. We

use the same different conditions as the ones described above for the shouted-
290 normal scenario:

- **All vs. All (A-A)**: All the whispered and normal speech utterances are verified against all the rest, which yields a total of 2,821,498 verification trials.
- **Normal vs. Normal (N-N)**: Normal speech utterances are verified
295 against the rest of neutrally-phonated utterances, which yields a total of 705,078 verification trials.
- **Whispered vs. Whispered (W-W)**: Whispered speech utterances are verified against the rest of whispered speech utterances, which yields a total of 705,078 verification trials.
- **Normal vs. Whispered (N-W)**: Normal speech utterances are verified
300 against whispered speech utterances, which yields a total of 1,411,342 verification trials.

For both databases, shouted-normal and whispered-normal, trials are generated automatically in order to compare each utterance in the dataset with
305 the rest. Afterwards, we divide the list of all possible pairs into several sublists to obtain the trials for all the different conditions explained before. These different trial lists are used to perform all speaker verification experiments. Finally, the verification scores, from which experimental results are calculated, are (computationally) obtained by comparing pairs of embeddings by means of
310 PLDA.

3.4. *Experimental Framework*

In this subsection, we explain the complete experimental framework used to perform the experiments. We have added to the baseline TDNN-PLDA speaker verification system three different modules: vocal effort detection, embedding
315 compensation and score calibration. In Figure 4, we can see the system diagram with the aforementioned three modules.

We trained the speaker verification system, TDNN and PLDA model, using augmented versions of VoxCeleb1 [40] and VoxCeleb2 [41]. We use this system to obtain the baseline results in the vocal effort conditions with the datasets explained in Subsection 3.3. Later on, we use the shouted-normal and whispered-normal (CHAINS) databases to apply the proposed vocal effort detection, embedding compensation and score calibration algorithms to improve baseline results. In Table 3, we describe the main characteristics of the databases used in this work.

Firstly, a vocal effort detector module is trained to classify the vocal effort of the embeddings extracted by the TDNN module. Two different detectors are trained depending on the database: one for normal-shouted and another for normal-whispered. The detectors, which are based on logistic regression, are trained and evaluated following a leave-one-speaker-out strategy by embeddings from the two datasets.

Then, the embedding compensation module is used to mitigate the vocal effort mismatch between embeddings. Experiments are divided into normal-shouted and normal-whispered embedding compensation, applied only to shouted and whispered embeddings (i.e., normal embeddings are not compensated). Experiments are also carried out by following a leave-one-speaker-out strategy.

Finally, the score calibration module is used to improve the system results at a score-level using the information about the vocal effort of the two embeddings that are compared. We separated the score calibration into two different experiments, normal-shouted and normal-whispered, training different logistic regression models with scores from different vocal effort conditions (see Section 5 for further details). Similarly as before, a leave-one-speaker-out strategy is used here to train and evaluate score calibration.

4. Vocal Effort Mode Detection

For a twofold reason, speaker verification systems are commonly trained using neutrally-phonated speech only: 1) this is the expected mode of use, and

Database	Purpose	Language	No. of Speakers	No. of Utterances
VoxCeleb1	Training	English	1,251	100,000
VoxCeleb2	Training	English	1,251	1,000,000
Normal-Shouted	Testing	Finnish	22	1,056
Normal-Whispered	Testing	English	36	2,376

Table 3: Main characteristics of the datasets used in our experiments.

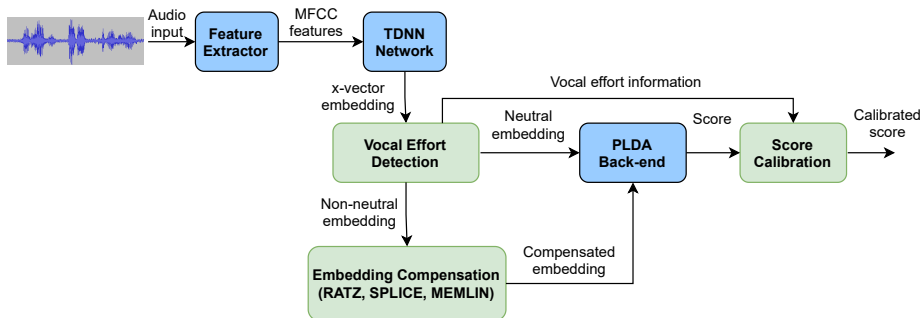


Figure 4: Block diagram of the speaker verification system used for experimental purposes in this paper including vocal effort detection, embedding compensation and score calibration modules.

2), as a result, much more data are available to train them. Hence, it makes sense to normalize the embeddings in terms of vocal effort in order to achieve a certain degree of robustness against vocal mode.

Let $\mathbf{z} = (z_1, \dots, z_D)^\top$ be a generic D -dimensional speaker embedding —recall that, in this paper, $D = 512$ (see Subsection 3.1)—. Furthermore, let $\mathbf{z}^{\{n\}}$, $\mathbf{z}^{\{w\}}$ and $\mathbf{z}^{\{s\}}$ represent speaker embeddings extracted from different utterances with the same phonetic content and by the same speaker in normal, whispered and shouted vocal modes, respectively. In this work, we are interested in functions $\mathbf{f}_{nw} : \mathbb{R}^D \rightarrow \mathbb{R}^D$ and $\mathbf{f}_{ns} : \mathbb{R}^D \rightarrow \mathbb{R}^D$ that map whispered and shouted speech embeddings, respectively, to equivalent normal speech embeddings, namely,

$$\begin{aligned}
 \mathbf{z}^{\{n\}} &= \mathbf{f}_{nw}(\mathbf{z}^{\{w\}}) \\
 &= \mathbf{f}_{ns}(\mathbf{z}^{\{s\}}).
 \end{aligned} \tag{1}$$

Evidently, applying these mapping functions to normal speech embeddings

350 $\mathbf{z}^{\{n\}}$ might yield distorted versions of them, $\tilde{\mathbf{z}}^{\{n_w\}} = \mathbf{f}_{n_w}(\mathbf{z}^{\{n\}})$ and $\tilde{\mathbf{z}}^{\{n_s\}} = \mathbf{f}_{n_s}(\mathbf{z}^{\{n\}})$, potentially harming the speaker verification system performance. To circumvent this issue, we perform vocal effort mode detection before applying this compensation. In this way, embeddings classified as neutral are not modified before scoring.

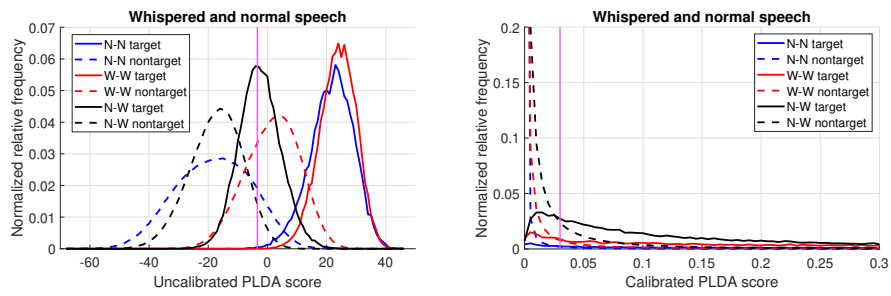
For vocal effort mode detection, we use logistic regression, since it is a low complexity and very effective approach for this purpose, as we showed in our previous work [1]. In particular, we train two independent logistic regression models for vocal mode classification: one to differentiate between normal and whispered speech embeddings, and another one to discriminate between normal and shouted embeddings. For the former case, let H_w and H_n be the hypotheses that \mathbf{z} is a whispered and a normal speech embedding, respectively. Therefore,

$$P(H_w|\mathbf{z}) = \frac{1}{1 + \exp\{-(\beta_{w,0} + \beta_{w,1}z_1 + \dots + \beta_{w,D}z_D)\}} \quad (2)$$

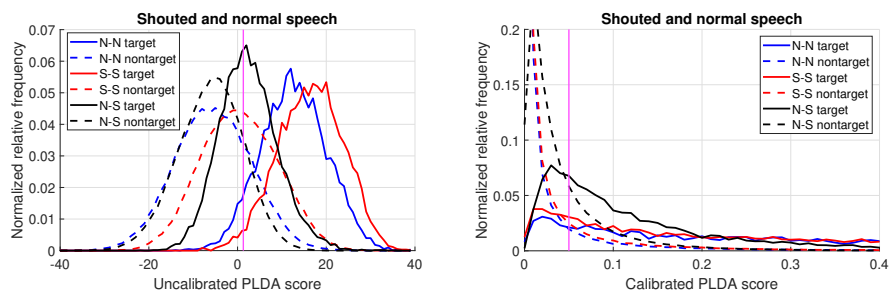
355 is the probability that the embedding \mathbf{z} was extracted from whispered speech. Notice that, on the contrary, the probability that \mathbf{z} was extracted from a neutrally-phonated utterance is $P(H_n|\mathbf{z}) = 1 - P(H_w|\mathbf{z})$. In (2), the set of parameters $\{\beta_{w,j}; j = 0, \dots, D\}$ is estimated from a set of whispered and normal speech training embeddings, which is derived from the speech data presented
 360 in Section 3. Particularly, estimation is carried out by minimizing binary cross-entropy with ℓ_2 regularization making use of the large-scale bound-constrained optimization algorithm L-BFGS-B [43]. Once the logistic regression model is trained, a new input embedding \mathbf{z} is considered to be a whispered speech embedding if $P(H_w|\mathbf{z}) > 0.5$ or normal otherwise. Notice that a parallel elaboration
 365 can be straightforwardly done for the logistic regression model discriminating between normal and shouted speech.

5. Score Calibration

It is also of interest to understand how different vocal effort modes impact the scores provided by speaker verification systems trained with normal speech



(a) Uncalibrated score distributions from whispered and normal speech comparisons. (b) Calibrated score distributions from whispered and normal speech comparisons.



(c) Uncalibrated score distributions from shouted and normal speech comparisons. (d) Calibrated score distributions from shouted and normal speech comparisons.

Figure 5: Score distributions from whispered (W) and normal (N) speech comparisons (top row), and from shouted (S) and normal (N) speech comparisons (bottom row). Left (right) column depicts uncalibrated (calibrated) score distributions. Calibrated scores were obtained by taking advantage of oracle vocal effort mode information. Equal error rate thresholds are indicated by magenta vertical lines. No vocal effort mode normalization was applied to x-vectors prior to PLDA scoring.

370 data. Towards this goal, Figures 5a and 5c show score distributions obtained
from whispered (W) and normal (N) speech embeddings, as well as from shouted
(S) and normal speech embeddings, respectively. It is worth noting that no vocal
effort mode normalization was applied to the embeddings before scoring. Ideally,
we would expect to observe two main, non-overlapping masses of probability,
375 that is, one for target scores and another for nontarget scores. However, this is
not what actually happens.

In Figure 5a, i.e., when dealing with both whispered and normal speech,
we can see a desirable behavior when performing N-N and W-W target scor-
ing, and N-N and N-W nontarget scoring, since the two latter distributions
380 highly overlap and are separated from the two former overlapping distribu-
tions. Nevertheless, the N-W target score distribution lies somewhere between
the aforementioned target and nontarget masses of probability due to the vo-
cal tract mismatch between whispered and normal modes. Similarly, the W-W
nontarget score distribution also lies between the two main target and nontarget
385 masses of probability. In this case, this could be due to the fact that vocal tract
characteristics across speakers experience a certain degree of standardization
under the whispered mode. An equivalent behavior can be observed in Figure
5c for the case of dealing with both shouted and normal speech.

In summary, Figures 5a and 5c reveal the adverse impact of vocal effort mode
390 mismatch on target and nontarget score discrimination, taking into account
that, commonly, speaker verification systems use a single score threshold to
decide whether or not a speech utterance corresponds to a claimed speaker.
Thus, we propose to exploit the information given by the logistic regression-
based vocal effort mode detector presented in Section 4. While, using such
395 information, one possibility consists of the definition of one score threshold per
vocal effort condition —i.e., N-N, W-W, N-W, S-S and N-S—, inspired by [44],
we perform logistic regression-based score calibration. Particularly, we train
one logistic regression model per comparison condition. In order to do that, we
train and evaluate the score calibration method with the leave-one-speaker-out
400 technique, using all the scores except the ones that correspond to the speaker

under evaluation to train the logistic regression models. Thanks to the vocal effort information, five different logistic regression models (N-N, W-W, N-W, S-S and N-S) are trained. In the evaluation phase, we obtain the calibrated scores using the corresponding logistic regression model according to the vocal effort condition of the involved utterances.

Figures 5b and 5d show the calibrated score distributions corresponding to Figures 5a and 5c, respectively. It is worth noting that these calibrated scores were obtained by taking advantage of oracle vocal effort mode information. While there is certainly a significant overlap between the target and nontarget masses of probability in Figures 5b and 5d, it can also be seen how the impact of the different comparison conditions has been compensated to a large extent. As a result, a more effective single score threshold can be set, improving the performance of the system from 23.54% EER in Figure 5a to 11.45% EER in Figure 5b, as well as from 24.76% EER in Figure 5c to 21.32% EER in Figure 5d.

In Section 7, we show that this score calibration methodology can be applied either standalone or in combination with the vocal effort mode compensation techniques of Section 6 for improved speaker verification performance.

6. Vocal Effort Compensation

In this section, vocal effort compensation methods based on [45], [46] and [47] are presented, where these techniques were originally proposed to mitigate the mismatch between noisy and clean speech features for automatic speech recognition purposes. In particular, these techniques, which are based on minimum mean square error (MMSE) estimation, are Multivariate Gaussian-based Cepstral Normalization (RATZ) [45], Stereo-based Piece-wise Linear Compensation for Environments (SPLICE) [46] and Multi-Environment Model-based Linear Normalization (MEMLIN) [47]. To employ these methods, it is first necessary to train different GMMs from MFCC features and learn a set of bias vectors to compensate the noisy MFCCs. In [1], these methods were used to

430 compensate speaker embeddings in order to mitigate the vocal effort mismatch
between normal and shouted speech.

From now on and for the sake of clarity, let \mathbf{x} be the normal speech embed-
ding corresponding to \mathbf{y} , which represents a speaker embedding extracted from
non-neutrally-phonated speech —i.e., from either shouted or whispered speech
435 in this particular work—. In addition, $\hat{\mathbf{x}}$ is an estimate of \mathbf{x} .

6.1. RATZ

In RATZ [45], the normal speech embedding space is modeled by means of
a GMM as follows:

$$p(\mathbf{x}) = \sum_{s_x} p(\mathbf{x}|s_x)P(s_x), \quad (3)$$

where $P(s_x)$ is the prior probability of the multivariate Gaussian s_x with mean
vector $\boldsymbol{\mu}_{s_x}$ and covariance matrix $\boldsymbol{\Sigma}_{s_x}$,

$$p(\mathbf{x}|s_x) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_{s_x}, \boldsymbol{\Sigma}_{s_x}). \quad (4)$$

Taking into consideration Eqs. (3) and (4), the RATZ estimator, $\mathbf{f}_{\text{RATZ}}(\mathbf{y}, s_x)$,
provides an estimate of \mathbf{x} as follows [45]:

$$\hat{\mathbf{x}} = \mathbf{f}_{\text{RATZ}}(\mathbf{y}, s_x) \approx \mathbf{y} - \sum_{s_x} \mathbf{r}_{s_x} \cdot p(s_x|\mathbf{y}), \quad (5)$$

where \mathbf{r}_{s_x} is a bias term that only depends on the normal speech Gaussian s_x
and is obtained in a previous training step with a set of paired embeddings from
both domains, $\{\mathbf{x}_i^{Tr}, \mathbf{y}_i^{Tr}\}$, according to

$$\mathbf{r}_{s_x} = \frac{\sum_i p(s_x | \mathbf{x}_i^{Tr}) (\mathbf{y}_i^{Tr} - \mathbf{x}_i^{Tr})}{\sum_i p(s_x | \mathbf{x}_i^{Tr})}. \quad (6)$$

6.2. SPLICE

The SPLICE method [46] is parallel to RATZ except for the former making
use of non-neutral vocal effort mode embeddings to train a GMM

$$p(\mathbf{y}) = \sum_{s_y} p(\mathbf{y}|s_y)P(s_y), \quad (7)$$

where, similarly, $P(s_y)$ is the prior probability of the multivariate Gaussian s_y with mean vector $\boldsymbol{\mu}_{s_y}$ and covariance matrix $\boldsymbol{\Sigma}_{s_y}$,

$$p(\mathbf{y}|s_y) = \mathcal{N}\left(\mathbf{y} \mid \boldsymbol{\mu}_{s_y}, \boldsymbol{\Sigma}_{s_y}\right). \quad (8)$$

Then, by considering Eqs. (7) and (8), the SPLICE estimator, $\mathbf{f}_{\text{SPLICE}}(\mathbf{y}, s_y)$, can be defined as follows in order to obtain an estimate of \mathbf{x} [46]:

$$\hat{\mathbf{x}} = \mathbf{f}_{\text{SPLICE}}(\mathbf{y}, s_y) \approx \mathbf{y} - \sum_{s_y} \mathbf{r}_{s_y} \cdot p(s_y|\mathbf{y}), \quad (9)$$

where \mathbf{r}_{s_y} is a bias term obtained in a training stage again using a set of paired embeddings from both domains, $\{\mathbf{x}_i^{\text{Tr}}, \mathbf{y}_i^{\text{Tr}}\}$, according to

$$\mathbf{r}_{s_y} = \frac{\sum_i p(s_y|\mathbf{y}_i^{\text{Tr}}) (\mathbf{y}_i^{\text{Tr}} - \mathbf{x}_i^{\text{Tr}})}{\sum_i p(s_y|\mathbf{y}_i^{\text{Tr}})}. \quad (10)$$

6.3. MEMLIN

MEMLIN [47] can be understood as a combination of RATZ and SPLICE, modeling both neutral and non-neutral speech domains by training two different GMMs in order to compensate the speaker embeddings derived from non-neutrally-phonated speech. Therefore, taking into account modeling of Eqs. (3) and (7), the MEMLIN estimator, $\mathbf{f}_{\text{MEMLIN}}(\mathbf{y}, s_x, s_y)$, can be expressed as [47]

$$\begin{aligned} \hat{\mathbf{x}} &= \mathbf{f}_{\text{MEMLIN}}(\mathbf{y}, s_x, s_y) \\ &= \int \sum_{s_y} \sum_{s_x} \mathbf{x} \cdot p(\mathbf{x}, s_x, s_y|\mathbf{y}) \cdot p(s_x, s_y|\mathbf{y}) d\mathbf{x} \\ &\approx \mathbf{y} - \sum_{s_y} \sum_{s_x} \mathbf{r}_{s_x s_y} \cdot p(s_y|\mathbf{y}) \cdot p(s_x|\mathbf{y}, s_y), \end{aligned} \quad (11)$$

where, similarly as before, $\mathbf{r}_{s_x s_y}$ is a bias term to be calculated in a training stage using a set of paired embeddings from both domains, $\{\mathbf{x}_i^{\text{Tr}}, \mathbf{y}_i^{\text{Tr}}\}$, in accordance with

$$\mathbf{r}_{s_x s_y} = \frac{\sum_i p(s_y, \mathbf{y}_i^{\text{Tr}}) p(s_x, \mathbf{x}_i^{\text{Tr}}) (\mathbf{y}_i^{\text{Tr}} - \mathbf{x}_i^{\text{Tr}})}{\sum_i p(s_y, \mathbf{y}_i^{\text{Tr}}) p(s_x, \mathbf{x}_i^{\text{Tr}})}. \quad (12)$$

In Figure 6a, we show normal and shouted speech x-vectors projected onto a two-dimensional space using t-SNE when MEMLIN is applied to shouted speech

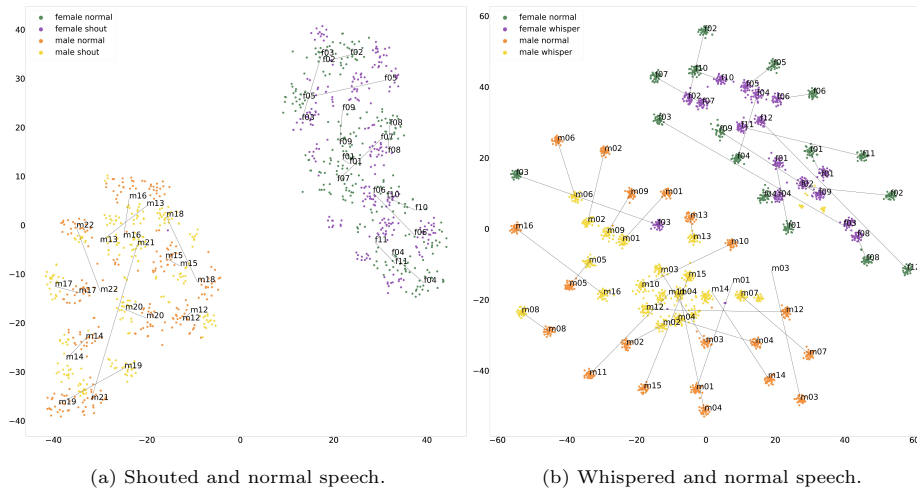


Figure 6: Speaker embedding representations obtained by means of t-SNE [42] when MEMLIN is employed for shouted and whispered speech embedding compensation.

embeddings. Unlike in Figure 3a, now, we can observe that both male and female speaker embeddings are grouped around their respective gender clusters regardless of the vocal effort condition. With this embedding compensation technique, it is possible to obtain embeddings from speech with different vocal effort modes in a same cluster, thereby improving the performance of the speaker verification system.

Furthermore, Figure 6b depicts normal and whispered speech embeddings projected onto a two-dimensional space, again using t-SNE, when also applying MEMLIN compensation to whispered speech x-vectors. We can see, in contrast to Figure 3b, how now, for each gender, whispered speech embeddings tend to overlay normal speech embeddings.

7. Results

In what follows, we present and discuss the results achieved with the proposed methods to detect and compensate the vocal effort mode of speaker embeddings, as well as to calibrate speaker verification scores.

Vocal Effort	Accuracy (%)	Vocal Effort Error (%)	Neutral Error (%)
Shouted	98.11	1.17	2.65
Whispered	99.88	0.00	0.11

Table 4: Shouted and whispered speech detection accuracy and error results, in percentages (%), when using logistic regression-based vocal effort detection.

7.1. Detection Results

The vocal effort detector is based on the algorithm explained in Section 4. For this task, two different detectors, namely, one for shouted speech and another one for whispered speech, have been trained and tested. The reason for
460 having two different detectors is that the datasets used to perform the experiments have different languages, speakers and recording conditions. Recall that, as mentioned in Subsection 3.3, due to data scarcity, we have made use of the leave-one-speaker-out strategy to train and test both vocal effort detectors.

According to the results shown in Table 4, the proposed shouted speech
465 detector obtains 98.11% accuracy with 1.17% error for shouted utterances and 2.65% error for neutrally-phonated utterances. On the other hand, the whispered speech detector also obtains very good results with 99.88% accuracy, and 0.11% and 0% error when detecting normal and whispered speech, respectively.

Despite the simplicity of this detection technique, it has demonstrated that
470 is able to obtain very accurate results in both databases, detecting shouted and whispered speech in a very appropriate way. For this reason, this is a key point to be applied before the embedding compensation and score calibration techniques.

7.2. Score Calibration Results

In this subsection, the results from applying the score calibration method
475 described in Section 5 are presented. Different models depending on each test database were trained and evaluated. Due to the lack of data, we again used a leave-one-speaker-out strategy.

In Table 5, we can see EER results *without* applying embedding compensa-
480 tion as well as using both oracle and logistic regression-based vocal effort mode
detection for shouted and whispered speech. All of the reported results in this
table belong to the All *vs.* All condition (see Section 3). In the first row, we
compare the baseline condition with the results obtained by applying score cal-
ibration from oracle detection of shouted speech. Here, we observe a 13.89%
485 relative improvement in terms of EER in comparison with the baseline.

In the second row, we can see the results using the proposed vocal effort
mode detection method for shouted speech. These results, as expected due to
the good performance of our logistic regression-based vocal effort detector, are
very similar to those obtained when applying oracle shouted speech detection.
490 Particularly, we achieve an 11.99% relative improvement in terms of EER with
respect to the baseline when considering score calibration.

In the third row, we show the result obtained with score calibration for
whispered speech and oracle vocal effort mode detection along with the baseline
result, which is achieved without score calibration. It can be seen that, only
495 using score calibration, we obtain a remarkable 51.35% relative improvement in
terms of EER in comparison with the baseline.

Finally, the last row of Table 5 presents the results of the same experiment
but using the proposed whispered speech detector. As it comes to no surprise,
we can observe almost the same numbers due to the fact that the whispered
500 speech detector is able to provide very good results (see Table 4).

7.3. Embedding Compensation Results

In this subsection, we show the results for the shouted-normal and whispered-
normal speech datasets comparing the baseline with the application of all the
methods described in Section 6 for mitigating vocal effort mismatch by means
505 of embedding compensation. Different experiments with the embedding com-
pensation methods (i.e., RATZ, SPLICE and MEMLIN) have been carried out.
We trained the GMMs used by these techniques making use of 2, 4, 8 and 16
components obtaining the best performance with 8 components. For the sake of

Vocal Effort	Detection Method	Baseline	Score Calibration
Shouted	Oracle	24.76	21.32
Shouted	Logistic regression	24.76	21.79
Whispered	Oracle	23.54	11.45
Whispered	Logistic regression	23.54	11.46

Table 5: Shouted-normal and whispered-normal speaker verification EER results, in percentages (%) and when using oracle and logistic regression-based vocal effort mode detection, from either applying (Score Calibration) or not (Baseline) score calibration. Notice that *no embedding compensation* is considered at this point.

Condition	Baseline	RATZ	SPLICE	MEMLIN
A-A	24.76	22.16	21.24	21.09
N-N	12.93	12.93	12.93	12.93
S-S	16.37	15.00	13.60	13.73
N-S	27.73	28.06	26.95	26.60

Table 6: Shouted-normal speaker verification EER results, in percentages (%) and when using oracle shouted speech detection, from either applying or not (Baseline) embedding compensation. Four different evaluation conditions are tested: A-A, N-N, S-S and N-S (see Section 3). Notice that *no score calibration* is considered at this point.

clarity, we decided not to include here all the results but only the most relevant
510 ones (i.e., those achieved by using 8-component GMMs). Furthermore, score calibration is *not* taken into account at this point. On the contrary, the combination of embedding compensation techniques with score calibration —applying oracle and logistic regression-based vocal effort detection— is left to the next subsection.

515 Table 6 shows embedding compensation experiments on the shouted-normal speech dataset that is arranged into four groups: All *vs.* All (A-A), Normal *vs.* Normal (N-N), Shouted *vs.* Shouted (S-S), and Normal *vs.* Shouted (N-S). In addition, different columns represent the baseline, and RATZ-, SPLICE- and MEMLIN-based embedding compensation. It is possible to appreciate

Condition	Baseline	RATZ	SPLICE	MEMLIN
A-A	24.76	22.26	21.34	21.62
N-N	12.93	13.14	13.20	12.93
S-S	16.37	15.15	13.76	14.03
N-S	27.73	28.17	27.00	26.43

Table 7: Shouted-normal speaker verification EER results, in percentages (%) and when using logistic regression-based shouted speech detection, from either applying or not (Baseline) embedding compensation. Four different evaluation conditions are tested: A-A, N-N, S-S and N-S (see Section 3). Notice that *no score calibration* is considered at this point.

520 that, under vocal effort mismatch conditions, EER increases substantially — the worst baseline scenario being N-S with an EER relative increase of around 114% with respect to N-N—. When embedding compensation techniques are used for shouted speech, better results are obtained. For example, MEMLIN achieves the largest EER relative improvements with respect to the baseline un-
525 der the A-A and N-S conditions, i.e., 14.82% and 4.07%, respectively. Moreover, for the S-S condition, SPLICE achieves the largest EER relative improvement in comparison with the baseline: 16.92%.

In Table 7, we show results from experiments equivalent to the ones above, where the only difference is that, now, logistic regression-based shouted speech
530 detection as in Section 4 is employed instead of oracle shouted speech detection. As expected, due to the reported performance of our vocal effort mode detector, we can observe very similar results to those reported in Table 6.

Next, we present the results obtained applying the same embedding compensation techniques but in the context of whispered-normal vocal effort by means
535 of the CHAINS database [17].

In Table 8, we present the results obtained applying the embedding compensation techniques RATZ, SPLICE and MEMLIN, by considering oracle whispered speech detection, in the following four different evaluation conditions: All *vs.* All (A-A), Normal *vs.* Normal (N-N), Whispered *vs.* Whispered (W-W),

Condition	Baseline	RATZ	SPLICE	MEMLIN
A-A	23.54	13.39	13.23	13.26
N-N	2.15	2.15	2.15	2.15
W-W	7.31	7.12	6.88	6.73
N-W	17.90	17.92	17.80	17.84

Table 8: Whispered-normal speaker verification EER results, in percentages (%) and when using oracle whispered speech detection, from either applying or not (Baseline) embedding compensation. Four different evaluation conditions are tested: A-A, N-N, W-W and N-W (see Section 3). Notice that *no score calibration* is considered at this point.

Condition	Baseline	RATZ	SPLICE	MEMLIN
A-A	23.54	13.39	13.58	13.27
N-N	2.15	2.15	2.15	2.17
W-W	7.31	7.12	7.60	6.73
N-W	17.90	17.92	18.18	17.85

Table 9: Whispered-normal speaker verification EER results, in percentages (%) and when using logistic regression-based whispered speech detection, from either applying or not (Baseline) embedding compensation. Four different evaluation conditions are tested: A-A, N-N, W-W and N-W (see Section 3). Notice that *no score calibration* is considered at this point.

540 and Normal *vs.* Whispered (N-W). We can observe that SPLICE provides the largest EER relative improvements with respect to the baseline under the A-A and N-W conditions, namely, 43.79% and 0.55%, respectively. Similarly, MEMLIN achieves a 7.93% EER relative improvement in the W-W scenario when compared to the baseline.

545 Table 9 presents experiments similar to the ones above, where the difference is that the proposed logistic regression-based vocal effort mode detector is used instead of oracle whispered speech detection. As we concluded for the shouted-normal scenario, due to the good results obtained by our logistic regression-based vocal effort detector, results in Table 9 are almost the same as those reported in Table 8.

550

Vocal Effort	Detection Method	Baseline	RATZ	SPLICE	MEMLIN
Shouted	Oracle	24.76	21.15	20.20	19.96
Shouted	Logistic regression	24.76	21.27	20.36	20.04
Whispered	Oracle	23.54	11.35	11.59	11.23
Whispered	Logistic regression	23.54	11.35	11.59	11.23

Table 10: Shouted-normal and whispered-normal speaker verification EER results, in percentages (%), when combining vocal effort mode detection (either oracle or based on logistic regression), embedding compensation and score calibration.

7.4. Embedding Compensation plus Score Calibration Results

In this subsection, we present the results obtained when we apply vocal effort mode detection (either oracle or based on logistic regression), embedding compensation and score calibration. It must be noted that all of the reported results in this subsection belong to the All *vs.* All condition.

As we can see from Table 10, combining the three methodologies described in this paper (i.e., vocal effort mode detection, embedding compensation and score calibration) provides the best results when dealing with either shouted or whispered speech in addition to neutrally-phonated speech. Endorsing the trend already seen in previous EER result tables, MEMLIN achieves the best performance among the different embedding compensation techniques tested. This can be attributed to the fact that MEMLIN, unlike RATZ and SPLICE, models both the neutral and non-neutral speech domains. Particularly, in the shouted-normal scenario, MEMLIN with score calibration achieves EER relative improvements of 19.38% and 19.06% with respect to the baseline when considering oracle and logistic regression-based vocal effort detection, respectively. For the whispered-normal condition, such EER relative improvements are 52.29% in both cases. Finally, let us notice that, compared to our previous work [1], we achieve a 7.3% relative improvement in terms of EER when adding score calibration to MEMLIN embedding compensation making use of our logistic regression-based vocal effort detector in the shouted-normal scenario.

8. Conclusions

In this work, we have studied the main differences between whispered, normal and shouted speech and how they affect to the performance of a speaker verification system that is trained with neutrally-phonated speech data. Due to the lack of data to train deep learning models with shouted and whispered speech data, we have developed different techniques to detect the vocal effort mode, perform score calibration and embedding compensation to mitigate vocal effort mismatch. Notable speaker verification performance improvements have been achieved in terms of EER from the combination of all of these techniques under vocal effort mismatch conditions, as we have experimentally shown.

According to the results, vocal effort-dependent calibration is crucial to improve the performance of the systems in situations where non-neutral speech is considered. In fact, compared to our previous work [1], we obtain a 7.3% relative improvement in terms of EER when adding score calibration in shouted-normal All *vs.* All condition. On the other hand, the best compensation technique to achieve the goals in this work is MEMLIN, showing similar improvements compared to SPLICE. The reason for that is mainly due to the fact that MEMLIN makes use of a greater number of parameters and therefore it presents greater modeling capabilities. MEMLIN is able to provide more accurate transformations by modeling both domains —i.e., normal and shouted/whispered speech—, while RATZ or SPLICE only model one of these spaces, normal or shouted/whispered speech. Depending on how we look at it, the relative simplicity of the compensation techniques studied in this work can be a strength or a weakness. In other words, while the modeling capabilities of these compensation techniques are very limited under a data rich scenario, they fit our vocal effort speech data scarcity conditions better than modern deep learning solutions that are typically data hungry. In fact, this data limitation can also be the explanation to the results in which SPLICE outperforms MEMLIN, since the former is simpler than the latter.

As future work, we will experiment with different models by exploring train-

ing data augmentation methodologies using different techniques in order to create speech samples under different vocal effort conditions for speaker verification system training purposes. In this way, we expect to substantially improve speaker verification performance without the need for embedding compensation and score calibration techniques.

9. Acknowledgements

This work was supported in part by the European Union’s Horizon 2020 research and innovation programme under Marie Skłodowska-Curie Grant 101007666; in part by MCIN/AEI/10.13039/501100011033 and by the European Union “NextGenerationEU” / PRTR under Grant PDC2021-120846-C41, and in part by the Government of Aragón (Grant Group T36_20R).

Authors would like to thank Dr. Tomi Kinnunen for providing the shouted speech corpus we have performed this study with and the researcher in Aalto University and University of Helsinki who made it possible [48].

References

- [1] S. Prieto, A. Ortega, I. López-Espejo, E. Lleida, Shouted speech compensation for speaker verification robust to vocal effort conditions, in: Proceedings of INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association, October 25-29, Shanghai, China, 2020, pp. 1511–1515.
- [2] M. Sahidullah, A. K. Sarkar, V. Vestman, X. Liu, R. Serizel, T. Kinnunen, Z.-H. Tan, E. Vincent, UIAI system for Short-Duration Speaker Verification Challenge 2020, in: Proceedings of SLT 2021 – IEEE Spoken Language Technology Workshop, January 19-22, Shenzhen, China, 2021, pp. 323–329.
- [3] Y. Jung, Y. Choi, H. Lim, H. Kim, A unified deep learning framework for short-duration speaker verification in adverse environments, *IEEE Access* 8 (2020) 175448–175466.

- 630 [4] A. Poddar, M. Sahidullah, G. Saha, Speaker verification with short utterances: a review of challenges, trends and opportunities, *IET Biometrics* 7 (2017) 91–101.
- [5] M. Sarria-Paja, T. H. Falk, Fusion of auditory inspired amplitude modulation spectrum and cepstral features for whispered and normal speech speaker verification, *Computer Speech and Language* 45 (2017) 437–456.
635
- [6] M. Sarria-Paja, T. Falk, Fusion of bottleneck, spectral and modulation spectral features for improved speaker verification of neutral and whispered speech, *Speech Communication* 102 (2018) 78–86.
- [7] A. R. Naini, A. R. M.V., P. K. Ghosh, Whisper to neutral mapping using cosine similarity maximization in i-vector space for speaker verification, in: *Proceedings of 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 4340–4344.
640
- [8] A. R. Naini, A. R. M. V., P. K. Ghosh, Formant-gaps features for speaker verification using whispered speech, in: *Proceedings of 44th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6231–6235.
645
- [9] C. Zhang, J. Hansen, Analysis and classification of speech mode: Whispered through shouted, in: *Proceedings of 8th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2007, pp. 2289–2292.
650
- [10] P. Zelinka, M. Sigmund, Automatic vocal effort detection for reliable speech recognition, in: *Proceedings of MLSP 2010 – IEEE International Workshop on Machine Learning for Signal Processing*, August 29–September 1, Kittilä, Finland, 2010, pp. 349–354.
- 655 [11] P. Zelinka, M. Sigmund, J. Schimmel, Impact of vocal effort variability on automatic speech recognition, *Speech Communication* 54 (2012) 732–742.

- [12] E. Jokinen, R. Saeidi, T. Kinnunen, P. Alku, Vocal effort compensation for MFCC feature extraction in a shouted versus normal speaker recognition task, *Computer Speech & Language* 53 (2019) 1–11.
- 660 [13] C. Hanilçi, T. Kinnunen, R. Saeidi, J. Pohjalainen, P. Alku, F. Ertaş, Speaker identification from shouted speech: Analysis and compensation, in: *Proceedings of 38th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8027–8031.
- [14] M. K. Nandwana, M. McLaren, L. Ferrer, D. Castan, A. Lawson, Analysis and mitigation of vocal effort variations in speaker recognition, in: *Proceedings of ICASSP 2019 – 44th International Conference on Acoustics, Speech and Signal Processing*, May 12-17, Brighton, UK, 2019, pp. 6001–6005.
- 665 [15] F. Kelly, J. H. Hansen, Analysis and calibration of Lombard effect and whisper for speaker recognition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021) 927–942.
- 670 [16] V. Vestman, D. Gowda, M. Sahidullaha, P. Alku, T. Kinnunen, Speaker recognition from whispered speech: A tutorial survey and an application of time-varying linear prediction, *Speech Communication* 99 (2018) 62–79.
- [17] T. L. F. Cummins, M. Grimaldi, J. Simko, The CHAINS corpus: Characterizing individual speakers, in: *Proceedings of SPECOM*, 2006, pp. 431–435.
- 675 [18] G. Klasmeyer, T. Johnstone, T. Bänziger, C. Sappok, K. R. Scherer, Emotional voice variability in speaker verification, in: *Proceedings of ITRW 2000 – ISCA Tutorial and Research Workshop on Speech and Emotion*, September 5-7, Newcastle, Northern Ireland, UK, 2000, pp. 213–218.
- 680 [19] S. Parthasarathy, C. Zhang, J. H. Hansen, C. Busso, A study of speaker verification performance with expressive speech, in: *Proceedings of ICASSP 2017 – 42nd International Conference on Acoustics, Speech and Signal Processing*, March 5-9, New Orleans, USA, 2017, pp. 5540–5544.

- 685 [20] B. H. Prasetyo, H. Tamura, K. Tanno, Emotional variability analysis based
i-vector for speaker verification in under-stress conditions, *MDPI Electronics* 9 (2020) 1–15.
- [21] B. D. Sarma, R. K. Das, Emotion invariant speaker embeddings for speaker
identification with emotional speech, in: *Proceedings of APSIPA ASC 2020*
690 – *Asia-Pacific Signal and Information Processing Association Annual Sum-*
mit and Conference, December 7-10, Auckland, New Zealand, 2020, pp.
610–615.
- [22] A. R. Avila, D. O’Shaughnessy, T. H. Falk, Automatic speaker verification
from affective speech using Gaussian mixture model based estimation of
695 neutral speech characteristics, *Speech Communication* 132 (2021) 21–31.
- [23] H. Chao, L. Dong, Y. Liu, Two-stage vocal effort detection based on spec-
tral information entropy for robust speech recognition, *Journal of Informa-*
tion Hiding and Multimedia Signal Processing 9 (2018) 1496–1505.
- [24] N. Obin, Cries and whispers - Classification of vocal effort in expressive
700 speech, in: *Proceedings of INTERSPEECH 2012 – 13th Annual Conference*
of the International Speech Communication Association, September 9-13,
Portland, USA, 2012, pp. 2234–2237.
- [25] P. Boersma, D. Weenink, Praat: doing phonetics by computer (2021).
URL <http://www.praat.org/>
- 705 [26] J.-S. Liénard, M.-G. D. Benedetto, Effect of vocal effort on spectral prop-
erties of vowels, *The Journal of the Acoustical Society of America* 106 (1999)
411–422.
- [27] J. Elliott, Comparing the acoustic properties of normal and shouted speech:
a study in forensic phonetics, in: *Proceedings of 8th Australian Interna-*
710 *tional Conference on Speech Science and Technology*, 2000, pp. 154–159.

- [28] P. Gramming, J. Sundberg, S. Ternström, R. Leanderson, W. H. Perkins, Relationship between changes in voice pitch and loudness, *Journal of Voice* 2 (1988) 118–126.
- [29] J. B. Wilson, J. D. Mosko, A comparative analysis of whispered and normally phonated speech using an LPC-10 vocoder, Tech. rep., Rome Air Development Center (1985).
715
- [30] S. T. Jovičić, Formant feature differences between whispered and voiced sustained vowels, *Acta Acustica united with Acustica* 84 (1998) 739–743.
- [31] T. Johnstone, The effect of emotion on voice production and speech acoustics, Ph.D. thesis, Psychology Department, University of Western Australia (2001).
720
- [32] G. Demenko, M. Jastrzębska, Analysis of natural speech under stress, *Acta Physica Polonica A* 121 (2012) 92–95.
- [33] M. V. Puyvelde, X. Neyt, F. McGlone, N. Pattyn, Voice stress analysis: A new framework for voice and effort in human performance, *Frontiers in Psychology* 9 (2018) 1–25.
725
- [34] J. H. Hansen, S. Patil, Speech under stress: Analysis, modeling and recognition, *Lecture Notes in Artificial Intelligence* 4343 (2007) 108–137.
- [35] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28 (1980) 357–366.
730
- [36] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. Lang, Phoneme recognition using time-delay neural networks, *IEEE Transactions on Acoustics, Speech, and Signal Processing* 37 (1989) 328–339.
735
- [37] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-vectors: Robust DNN embeddings for speaker recognition, in: *Proceedings*

of 43rd International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5329–5333.

- 740 [38] Z. Bai, X.-L. Zhang, Speaker recognition based on deep learning: An overview, arXiv:2012.00931v2 (2021).
- [39] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, The Kaldi speech recognition toolkit, in: Proceedings of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, IEEE Signal Processing Society, 2011, iEEE Catalog No.: CFP11SRW-USB.
- 745 [40] A. Nagrani, J. S. Chung, A. Zisserman, VoxCeleb: a large-scale speaker identification dataset, in: Proceedings of 18th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2017, pp. 2616–2620.
- [41] J. S. Chung, A. Nagrani, A. Zisserman, VoxCeleb2: Deep speaker recognition, in: Proceedings of 19th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2018, pp. 1086–1090.
- 750 [42] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, *Journal of Machine Learning Research* 9 (2008) 2579–2605.
- [43] J. L. Morales, J. Nocedal, Remark on “algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound constrained optimization”, *ACM Transactions on Mathematical Software* 38 (2011) 1–4.
- 760 [44] F. Kelly, J. H. L. Hansen, Detection and calibration of whisper for speaker recognition, in: Proceedings of SLT 2018 – IEEE Spoken Language Technology Workshop, December 18-21, Athens, Greece, 2018, pp. 1060–1065.
- [45] P. J. Moreno, B. Raj, E. B. Gouvêa, R. M. Stern, Multivariate-Gaussian-based cepstral normalization for robust speech recognition, in: Proceedings

- 765 of 1995 International Conference on Acoustics, Speech, and Signal Process-
ing, ICASSP '95, Detroit, Michigan, USA, May 8-12, 1995, IEEE Computer
Society, 1995, pp. 137–140. doi:10.1109/ICASSP.1995.479292.
URL <https://doi.org/10.1109/ICASSP.1995.479292>
- [46] J. Droppo, L. Deng, A. Acero, Evaluation of the SPLICE algorithm on the
770 Aurora2 database, in: Proceedings of 7th European Conference on Speech
Communication and Technology (EUROSPEECH), 2001, pp. 217–220.
- [47] L. Buera, E. Lleida, A. Miguel, A. Ortega, O. Saz, Cepstral vector normal-
ization based on stereo data for robust speech recognition, IEEE Trans-
actions on Audio, Speech, and Language Processing 15 (2007) 1098–1113.
775 doi:10.1109/TASL.2006.885244.
- [48] T. Raitio, A. Suni, J. Pohjalainen, M. Airaksinen, M. Vainio, P. Alku, Anal-
ysis and synthesis of shouted speech, in: Proceedings of INTERSPEECH
2013 – 14th Annual Conference of the International Speech Communication
Association, August 25-29, Lyon, France, 2013, pp. 1544–1548.



Click here to access/download

LaTeX Source Files
WN_Baseline.eps





Click here to access/download
LaTeX Source Files
Whispered.eps

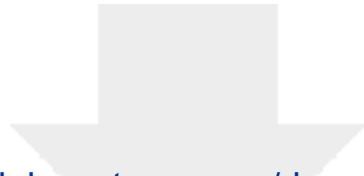




Click here to access/download

LaTeX Source Files
SN_Baseline.eps

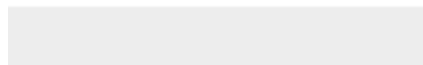




Click here to access/download

LaTeX Source Files

XVECTORS_SOLO_WHISPER_BASELINE.pdf

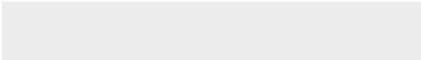


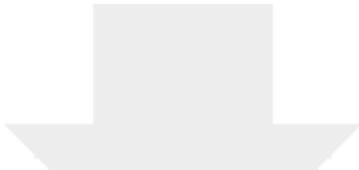


Click here to access/download

LaTeX Source Files

[WN_Baseline-eps-converted-to.pdf](#)

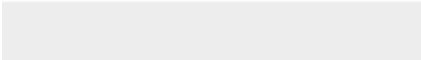




Click here to access/download

LaTeX Source Files

[SN_Calibrated-eps-converted-to.pdf](#)





Click here to access/download
LaTeX Source Files
Shouted.eps





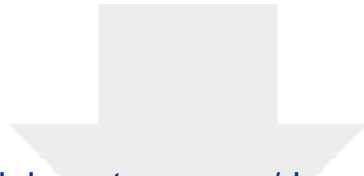
Click here to access/download
LaTeX Source Files
Legend.eps





Click here to access/download
LaTeX Source Files
elsarticle-num.bst

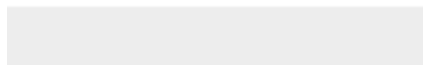
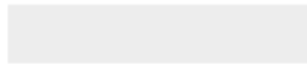




Click here to access/download

LaTeX Source Files


XVECTORS_SOLO_WHISPER_MEMLIN.pdf





Click here to access/download
LaTeX Source Files
elsarticle.dtx



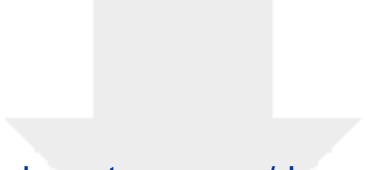


Click here to access/download

LaTeX Source Files

[XVECTORS_NORMAL_SHOUTED_MEMLIN.pdf](#)



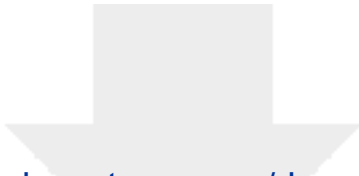


Click here to access/download

LaTeX Source Files

XVECTORS_NORMAL_SHOUTED_BASELINE.pdf

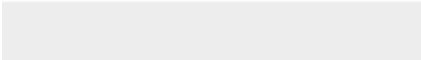




Click here to access/download

LaTeX Source Files

[WN_Calibrated-eps-converted-to.pdf](#)





Click here to access/download
LaTeX Source Files
WN_Calibrated.eps

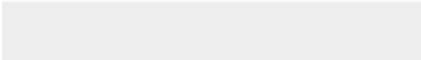





Click here to access/download

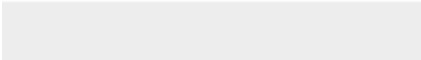

LaTeX Source Files

Whispered-eps-converted-to.pdf





Click here to access/download
LaTeX Source Files
svs_diagram_new.pdf





Click here to access/download
LaTeX Source Files
svs_diagram.pdf





Click here to access/download

LaTeX Source Files
SN_Calibrated.eps

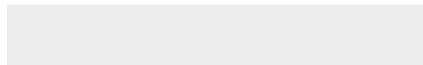


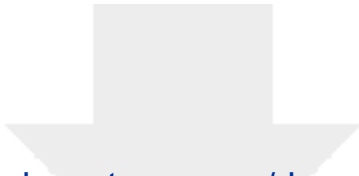


Click here to access/download

LaTeX Source Files

[SN_Baseline-eps-converted-to.pdf](#)

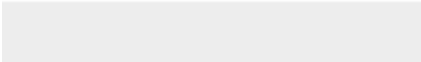




Click here to access/download

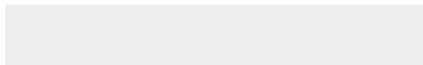
LaTeX Source Files

Shouted-eps-converted-to.pdf





Click here to access/download
LaTeX Source Files
Normal-eps-converted-to.pdf



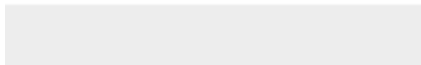


Click here to access/download
LaTeX Source Files
Normal.eps





Click here to access/download
LaTeX Source Files
Legend-eps-converted-to.pdf





Click here to access/download
LaTeX Source Files
mybib.bib





Click here to access/download

LaTeX Source Files
Paper.tex



Santi Prieto: Software, Validation, Investigation, Data Curation, Writing - Original Draft, Visualization. **Alfonso Ortega:** Conceptualization, Methodology, Formal analysis, Resources, Supervision. **Iván-López-Espejo:** Writing - Review & Editing, Visualization. **Eduardo LLeida:** Supervision, Writing - Review & Editing.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: