Alku, Paavo; Kodali, Manila; Laaksonen, Laura; Kadiri, Sudarsana

AVID: A speech database for machine learning studies on vocal intensity

Published in:
Speech Communication

DOI:
10.1016/j.specom.2024.103039

Published: 01/02/2024

Document Version
Publisher's PDF, also known as Version of record

Speech Communication 157 (2024) 103039

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

AVID: A speech database for machine learning studies on vocal intensity

Paavo Alku [a], Manila Kodali [a,*], Laura Laaksonen [b], Sudarsana Reddy Kadiri [a]

[a] Department of Information and Communications Engineering, Aalto University, Finland
[b] Tampere Wireless Headset Audio Lab, Huawei Technologies Oy Co., Ltd., Finland

ARTICLE INFO

Dataset link: https://zenodo.org/doi/10.5281/zenodo.7948299

Keywords:
Vocal intensity
Speech database
Sound pressure level
Support vector machine
Convolutional neural network
Machine learning

ABSTRACT

Vocal intensity, which is quantified typically with the sound pressure level (SPL), is a key feature of speech. To measure SPL from speech recordings, a standard calibration tone (with a reference SPL of 94 dB or 114 dB) needs to be recorded together with speech. However, most of the popular databases that are used in areas such as speech and speaker recognition have been recorded without calibration information by expressing speech on arbitrary amplitude scales. Therefore, information about vocal intensity of the recorded speech, including SPL, is lost. In the current study, we introduce a new open and calibrated speech/electroglottography (EGG) database named Aalto Vocal Intensity Database (AVID). AVID includes speech and EGG produced by 50 speakers (25 males, 25 females) who varied their vocal intensity in four categories (soft, normal, loud and very loud). Recordings were conducted using a constant mouth-to-microphone distance and by recording a calibration tone. The speech data was labelled sentence-wise using a total of 19 labels that support the utilisation of the data in machine learning (ML) -based studies of vocal intensity based on supervised learning. In order to demonstrate how the AVID data can be used to study vocal intensity, we investigated one multi-class classification task (classification of speech into soft, normal, loud and very loud intensity classes) and one regression task (prediction of SPL of speech). In both tasks, we deliberately warped the level of the input speech by normalising the signal to have its maximum amplitude equal to 1.0, that is, we simulated a scenario that is prevalent in current speech databases. The results show that using the spectrogram feature with the support vector machine classifier gave an accuracy of 82% in the multi-class classification of the vocal intensity category. In the prediction of SPL, using the spectrogram feature with the support vector regressor gave an mean absolute error of about 2 dB and a coefficient of determination of 92%. We welcome researchers interested in classification and regression problems to utilise AVID in the study of vocal intensity, and we hope that the current results could serve as baselines for future ML studies on the topic.

1. Introduction

In speech communication, speakers regulate vocal intensity on many occasions, for example, to emphasise something, to be heard in noisy conditions or over a long distance, or to signal vocal emotions such as anger or sadness. Vocal intensity is quantified typically with the sound pressure level (SPL), which is defined in the dB scale as the logarithm of the ratio between the sound pressure and the standard reference pressure of 20 μPa (see Eq. 9.2b in Titze (1994)). In their tutorial on measurement of SPL (Švec and Granqvist, 2018), Švec and Granqvist described two approaches that can be used in speech and voice recordings to measure SPL. The first approach is to use a special device, a sound level meter (SLM), to register SPL. The second approach corresponds to first recording a standard calibration tone (a sinusoidal of 1 kHz with a reference SPL of 94 dB or 114 dB) and then computing SPL as the RMS (root mean square) ratio between the speech signal and the calibration tone (Švec and Granqvist, 2018). The latter approach is beneficial because it can be done with cheaper equipment (i.e. a calibrator) and it enables computing SPL values from the recorded speech signals afterwards, which in turn, makes it possible to vary settings such as the frequency weighting or the speech unit over which the SPL is computed. In both approaches, the mouth-to-microphone distance is essential and should always be reported with SPL measurements because the obtained SPL values are higher when measured close to the mouth and lower when captured further away from the mouth (Švec and Granqvist, 2018). In the past few decades, SPL measurements have been reported in many studies investigating intensity regulation of speech (e.g. Holmberg et al., 1988; Titze and Sundberg, 1992; Hodge et al., 2001; Alku et al., 2006, 2002; Liénard, 2019; Nash, 2014), and these studies have shown that healthy speakers can vary vocal intensity over a wide SPL range. In an investigation

* Corresponding author.
E-mail address: manila.kodali@aalto.fi (M. Kodali).

https://doi.org/10.1016/j.specom.2024.103039
Received 14 June 2023; Received in revised form 31 October 2023; Accepted 16 January 2024
Available online 23 January 2024

by Coleman et al. (1977), for example, speakers produced vowels using their maximum and minimum vocal effort, and the results showed that the SPL values (measured using a mouth-to-microphone distance of 6 in.) varied from 48 dB in soft speech to 126 dB in very loud speech.

Audio equipment (e.g. loudspeakers) change sound intensity by simply amplifying or attenuating the level of the sound waveform. In other words, increasing intensity in audio equipment corresponds to multiplying the sound waveform with a constant that is larger than unity and decreasing intensity corresponds to multiplying the waveform with a constant that is smaller than unity. The intensity regulation of natural speech is conducted by the physiological human voice production apparatus, and it uses various complicated mechanisms. As discussed by Titze in Titze (1994), the intensity regulation mechanisms of natural speech consist of three parts corresponding to adjustments below the larynx, within the larynx and above the larynx. Below the larynx, intensity is affected by controlling the aerodynamic output of the lungs (particularly by varying the subglottal pressure) to the vocal system. Within the larynx, regulation of intensity happens by modifying the vibration of the vocal folds. An increase in intensity is obtained by raising the flow amplitude through the glottis and by decreasing the length of the glottal closing phase. Above the larynx, vocal intensity can be modified by adjusting the resonances of the vocal cavity, especially the first formant, to coincide with the harmonics of the glottal source. Due to the complex nature of the above-described intensity regulation mechanisms in voice production, changing the intensity of natural speech is not a simple multiplication of the waveform with a constant gain as in audio equipment, but the produced speech changes in many ways in terms of its acoustic and prosodic characteristics. Previous studies have shown that the fundamental frequency (F0) and the first formant of speech typically rise and the spectral tilt lowers when vocal intensity is raised (e.g. Hodge et al., 2001; Huber et al., 1999; Lienard and Benedetto, 1999). In terms of prosodic features, previous studies have shown that raising of vocal intensity affects phone durations shortening bilabial stops and lengthening stressed vowels (Schulman, 1989).

Even though regulation of vocal intensity is an important topic in speech communication, speech databases that are widely used today in speech technology (e.g. TIMIT (Anon, 1993), LibriSpeech (Anon, 2022), NIST SRE (Greenberg et al., 2020)) do not include information about SPL. In other words, speech signals of these databases have been collected without using a constant mouth-to-microphone distance combined with either registering SPL during the recordings using an SLM or by recording a (standard) calibration tone to measure SPL from the recorded speech afterwards. We argue that there are two potential explanations for the absence of SPL information in most current speech databases. First, many popular openly available speech databases have been collected for the purposes of data-driven speech technology topics such as automatic speech recognition (ASR), speaker recognition (SR) and text-to-speech synthesis (TTS) that call for expertise in machine learning (ML). It might be that even though SPL measurements are in principle easy to conduct, many researchers in these popular speech technology domains are simply not as familiar with acoustical measurement of SPL as researchers in the voice research community. Second, current speech databases have been collected mainly to provide training and testing data to develop ML models for popular research areas, particularly ASR, SR and TTS. For these areas, SPL of the recorded speech is not of interest because training/testing of ML models can be done using time domain speech signals represented on arbitrary amplitude scales (e.g. by scaling the recorded speech signal so that its maximum absolute value is 1.0).

In the current article, we introduce a new open database collected at Aalto University (Finland) that includes speech signals produced in different vocal intensity categories.[1] The database, called the Aalto Vocal Intensity Database (AVID), aims to help research of both the speech technology community and the voice research community by providing open access speech data to investigate vocal intensity. The launching of AVID is motivated by two goals. First, AVID aims to contribute to basic research in speech and voice production by providing open access to speech and electroglottography (EGG) data that has been produced in four different intensity categories (soft, normal, loud and very loud) and recorded together with a standard calibration tone using a constant mouth-to-microphone distance. Second, since the SPL values of the recorded speech signals can be determined (due to the calibrated recordings), it is possible to utilise the AVID data to study the following new interesting research question: By using the intensity category or the SPL of a recorded speech signal as ground truth, would it be possible to train ML networks to predict the signal's intensity category or its SPL even though the signal is presented using an arbitrary amplitude scale as in most current databases? Studying this problem is feasible because speech includes acoustical cues brought about by the human intensity regulation mechanism, and these cues could be used by the machine despite the most obvious acoustical intensity cue (i.e. information about the signal's amplitude) is absent. Two interesting ML tasks can be studied in this research area: (1) prediction of the intensity category of a speech signal, which is a multi-class classification problem and (2) prediction of the SPL of a speech signal, which is a regression problem. The results achieved by studying these tasks can be applied in speech-based biomarking of health and in forensic speech research. In the former, the prediction of the vocal intensity category could be used in the detection of disorders such as Parkinson's disease and heart failure that have shown to affect the intensity regulation mechanisms of speech (Clark et al., 2014; Fox and Ramig, 1997; DeKeyser et al., 2016; Mittapalle et al., 2022). In the latter application area, automatic classification of vocal intensity category is of interest to distinguish speech recordings where the speaker has deliberately used soft speech in order not to be overheard (Zhang and Hansen, 2011). Previous studies have mainly addressed binary classification problems in which ML has been used to automatically distinguish a certain vocal intensity mode such as whispering (e.g. Zhang and Hansen, 2011; Meenakshi and Ghosh, 2015; Sarria-Paja and Falk, 2013) or shouting (e.g. Pohjalainen et al., 2013; Baghel et al., 2020; Laffitte et al., 2016) from speech of normal loudness. However, there are only a few investigations (Kodali et al., 2023b; Zhang and Hansen, 2007; Zelinka et al., 2012) that have addressed the multi-class classification problem and to the best of our knowledge, the regression problem has not been studied yet in any article. We argue that the best way to raise general interest in the proposed ML-based research questions is to use the strategy that has been followed in many recent works (Sharma et al., 2021; Zhou et al., 2022; Wielgat et al., 2021; Alku et al., 2019; Kibria et al., 2022), that is, to publish a journal article that describes the background and details of the speech corpus collected.

This article is organised as follows. In Section 2, the recording of the AVID database is described together with the organisation of the database and labelling of the data. Section 3 visualises the collected data in terms of three acoustical features. In Section 4, the ML experiments to be conducted in the study are explained. Section 5 provides the results of the ML experiments by first reporting the results of the multi-class classification experiments and then the results of the regression experiments. Finally, a discussion and conclusions of the study are provided in Sections 6 and 7, respectively.

[1] In this study, we use the term 'vocal intensity', which is widely used in speech acoustics and voice research (e.g. Titze, 1994). In phonetics, the term 'vocal effort' is used in studying the same phenomena (Traunmüller and Eriksson, 2000). We regard these two terms as synonymous.

Fig. 1. A speaker conducting a speaking task in the recording environment.

2. Data

2.1. Recording

Speech signals were recorded from 50 speakers (25 males and 25 females), who produced speech in four intensity categories (soft, normal, loud, and very loud). In each category, the speakers were allowed to use their habitual regulation of vocal intensity. In addition to the acoustic speech signal, we also recorded electroglottography (EGG) signals that measure the degree of contact between the vocal folds. (For a review of EGG, see Herbst (2020)). Though both the acoustic speech signal and the EGG signal were collected, this article focuses solely on the former because it is the key modality in speech technology. All the speakers were students at Aalto University, and none of them had a history of any speech, voice or hearing disorders. The male speakers were 20 to 38 years old, and the female speakers were 21 to 31 years old. The speech signals were produced in English, and all participants were proficient in English. The recordings were performed in a listening room that fulfilled the requirements of ITU-R BS.1116-1 (Rec, 1997). The DPA 4065-BL headset condenser microphone was used to record the acoustic speech signal. The microphone was placed at a distance of 5 cm from the speaker's lips. To record the EGG signal, the EG2-PCX2 EGG device was used. The recordings were conducted using a sampling frequency of 44.1 kHz and a resolution of 32 bits. The other equipment used in the recordings were an Amprobe SM-CAL1 calibrator (Anon, 2021a), an RME Babyface sound card (Anon, 2021b) and a laptop. Using the calibrator, a calibration tone (a sinusoidal of 1 kHz) of 94 dB in SPL was recorded before each speaker's speech/EGG recording.

To elicit soft speech, the speaker was instructed to 'Speak softly but do not whisper; speak as you would talk to your peer in a lecture.' For normal speech, the speaker was instructed to 'Speak as you would talk to your friend during a lecture break and intervals.' In order to elicit loud speech, the speaker was instructed to 'Speak as a lecturer,' and for very loud speech to 'Speak as you would talk to someone in a noisy room but do not to shout.' Note that the speaker was not instructed to use a predefined SPL in any category but he/she was allowed to use his/her habitual vocal intensity in each four category. Fig. 1 shows a speaker in the recording environment.

The data collection process was divided into two sessions (Session 1 and Session 2) both of which consisted of two speaking tasks: the sentence reading task (SENT) and the paragraph reading task (PARA). Session 2 was a repetition of Session 1, and the tasks were identical in both sessions. In the SENT task, each speaker was asked to recite 25 isolated sentences in each of the four intensity categories. The orthographic transcriptions of the sentences were selected from the TIMIT database (Garofolo, 1993). To make the speaker produce sentences of different duration, orthographic transcriptions of different lengths were selected from TIMIT by varying the number of words per sentence from three to seven. This procedure yielded five sentence lengths, and by selecting five different orthographic transcriptions for each sentence length, 25 sentences were obtained in each intensity category. In the PARA task, each speaker was asked to recite two different text paragraphs in all four vocal intensity categories. The first paragraph was identical for all the speakers, and it was taken from a weather forecast (Anon, 2021c). The second paragraph was different for the speakers, and it was obtained by selecting excerpts from a novel (Anon, 2008) that covered approximately an equal number of words for each speaker. The size of the raw data in AVID is approximately 16 h. In total, the SENT task consisted of 10 000 sound files (25 sentences × 50 speakers × 4 intensity categories × 2 sessions), with 2500 files per each intensity category. The PARA task consisted of 800 sound files (2 paragraphs × 50 speakers × 4 intensity categories × 2 sessions), with 200 files per each intensity category. Table 1 shows the number of speech files in both the SENT and PARA tasks for each of the four intensity categories.

Table 1
Number of speech files in each of the four intensity categories for the two speaking tasks (SENT and PARA) of the AVID database.

Intensity category	SENT task	PARA task
Soft	2500	200
Normal	2500	200
Loud	2500	200
Very loud	2500	200
Total	10 000	800

2.2. Organisation of AVID

The collected data was saved to the AVID database, and the repository was organised using the structure shown in Fig. 2. The recorded raw data, including both the speech and EGG waveforms, was saved to Repository 1 as stereo wav files using the original sampling frequency of 44.1 kHz. Since the main goal of AVID is to provide training and testing data for speech-based ML experiments in the study of vocal intensity, the acoustic speech signals were separated from the original stereo recordings and saved to Repository 2 as mono wav files.

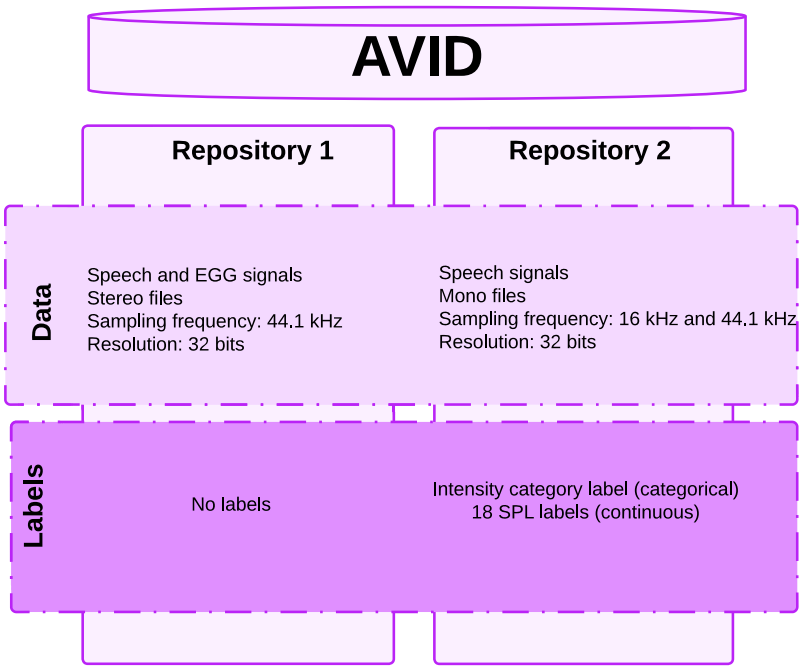

Fig. 2. Repository structure of the AVID database. The recorded raw data (speech and EGG) was saved to Repository 1. Repository 2 consists of manually segmented one-sentence-long labelled speech signals.

Repository 2 was divided into two sections: the first one for the speech signals downsampled to 16 kHz, and the second one for the speech signals saved using the original sampling frequency of 44.1 kHz. In this study, we only use the signals of the former section of Repository 2, because 16 kHz is the *de facto* sampling frequency for speech in ML experiments. All the speech signals saved into Repository 2 are one-sentence-long signals that were manually segmented from the recorded signals produced both in the SENT and PARA tasks.

2.3. Labelling: Intensity category labels and SPL labels

In order to use the collected speech data in multi-class classification studies of vocal intensity, the sentences of Repository 2 were labelled with the intensity category (i.e. soft/normal/loud/very loud) used by the speaker in the production of the corresponding signal. In the remainder of this article, we refer to this categorical label as the '*intensity category label*'. We would like to point out that with this label, the recorded sentences are separated into four distinct classes in the most straightforward manner according to the speaking task given to the speaker. It is worth reminding that speakers were allowed to use their habitual regulation of vocal intensity in the recording and therefore it is possible that a sentence labelled 'soft' has a larger SPL than a sentence labelled 'normal' because the former was produced by a speaker that simply has a habitually loud voice. Hence, the intensity category label is not based on the SPL of the produced sentence but is *subjective*.

Instead of labelling the collected data using only the straightforward approach described above, we supplied the AVID database with a rich set of objective SPL values, called the '*SPL labels*', in order to use the data in a more diverse range of classification and regression studies of vocal intensity. As will be demonstrated in Section 4, these continuous SPL labels can be used in two types of ML experiments: (1) in automatic classification of speech into intensity classes that are defined objectively using SPL boundaries and in (2) regression studies, where the SPL of a speech sentence is predicted. The AVID database provides each sentence of Repository 2 with as many as 18 SPL labels that cover the most prevalent SPL parameters in frequency weighting, time weighting and time averaging, as described in Švec and Granqvist (2018). The SPL labels were computed as follows. First, we removed silent segments and segments of unvoiced speech from the recorded signals, as recommended in Švec and Granqvist (2018), Švec et al. (2005). The removal of silence was conducted using the SOX tool (Barras, 2012), which assumes that frames of active speech can be separated from silence based on an energy threshold. Second, by using the silence-removed speech signals together with the calibration tone, we computed 18 different SPL values for each sentence in Repository 2 by varying the following commonly used SPL settings: frequency weighting, time weighting and time averaging. For frequency weighting, we used the A-, C-, and Z (i.e., zero) weightings. For time weighting, we used the slow (S), medium (M) and fast (F) weightings with the time constant τ of 1 s, 0.125 s and 0.03 s, respectively. For time averaging, we used the mean SPL and the equivalent SPL. Hence, by combining three frequency weightings, three time weightings and two time averaging methods, each sentence was labelled with a total of 18 different SPL values. All these computations were conducted using the software tool available in Švec and Granqvist (2018). More details of the frequency weightings, time weightings and time averaging methods can be found in Švec and Granqvist (2018).

In the remainder of this article, we refer to the SPL of a speech signal using a notation that specifies how the settings described above were selected. We use the notation L_{ijk}, where L is the SPL measured in decibels, and i, j, and k refer to the time averaging (equivalent (eq) or mean), frequency weighting (A, C or Z) and time weighting used (S, M or F), respectively. As an example, L_{eqZF} refers to the SPL value measured using the equivalent time averaging, Z frequency weighting and fast time weighting.

3. Acoustical analyses and visualisation of the AVID speech data

The speech signals from the SENT task were analysed between the different vocal intensity categories using three parameters: SPL (L_{eqZF}), fundamental frequency (F0) and spectral tilt. Spectral tilt was estimated using the first mel-frequency cepstral coefficient (MFCC-1). These three parameters were first computed for each individual sentence and then averaged across all the sentences spoken by each speaker. The parameters were expressed using violin plots to visualise and compare

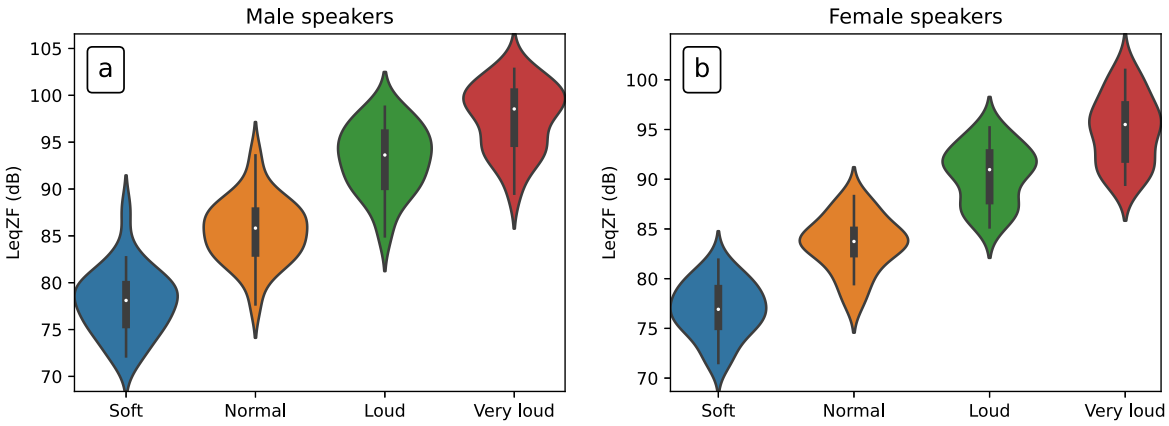

Fig. 3. Distribution of SPL(L_{eqZF}) for male (a) and female (b) speakers in the four intensity categories used in the recording of AVID.

the parameter distributions between the different intensity categories. Fig. 3 shows the distribution of SPL (L_{eqZF}) for each vocal intensity category separately for male and female speakers. The figure shows an increasing trend from soft to very loud, with the lowest SPL values for soft and the highest SPL values for very loud. Similarly, the distribution of F0, depicted in Fig. 4, shows that the fundamental frequency is highest in very loud phonation and lowest in soft phonation, with normal and loud phonations showing intermediate values. Fig. 5 shows that the speakers' spectral tilt values decrease from the soft mode to the very loud mode (i.e. spectral envelopes become flatter).

To further analyse the three parameters, one-way ANOVA tests were performed to examine statistical differences in the parameter values between the intensity categories. The ANOVA tests were conducted separately for each of the three parameters, treating the underlying parameter as a dependent variable and the intensity category as an independent variable. The null hypothesis was that there was no difference in the parameter values between the vocal intensity categories. The results showed that the intensity category had a significant effect on all three parameters (SPL in females: $F(3, 96) = 176.0$, p-value $<$ 1.0e−38; SPL in males: $F(3, 96) = 152.0$, p-value $<$ 1.0e−35; F0 in females: $F(3, 96) = 26.1$, p-value $<$ 1.0e−11; F0 in males: $F(3, 96) = 27.0$, p-value $<$ 1.0e−12; MFCC-1 in females: $F(3, 96) = 64.0$, p-value $<$ 1.0e−23; MFCC-1 in males: $F(3, 96) = 35.0$, p-value $<$ 1.0e−15).

The distributions shown in Figs. 3–5 indicate statistically significant results, which suggest that the speakers performed their speaking tasks properly. Furthermore, the results demonstrate significant differences between the intensity classes, not only in the main measure of vocal intensity (SPL) but also in two other acoustical speech features (F0 and spectral tilt). Therefore, the collected material is well-suited to be used as training and testing data in ML-based classification and regression studies of vocal intensity.

4. ML experiments

As explained in Section 1, one motivation for the current study is to raise awareness of the speech and voice research communities for ML-based studies of vocal intensity. We are particularly advocating the utilisation of ML in a scenario where the original intensity information of speech is lost because the signal has been recorded without SPL calibration and is therefore expressed on an arbitrary amplitude scale. In order to demonstrate how ML models can be used together with the AVID database for these kinds of research problems, the following two tasks are studied in this section: (1) classification of the intensity category of speech, which is a multi-class classification task and (2) prediction of SPL of speech, which is a regression task. A schematic diagram describing these two ML tasks is shown in Fig. 6. This sub-section describes the components of the pipeline used for two tasks, including labels, features and models employed, and evaluation metrics. All the experiments described in this section are based on the sentence data collected in Repository 2. Speech is expressed according to the scenario advocated above, that is, the original intensity information is deliberately removed by normalising the maximum amplitude of each time-domain signal in each intensity category to 1.0.

Table 2
Division of the speech signals into four SPL-based intensity categories based on SPL boundaries (type of SPL: L_{meanZF}).

SPL-based intensity category	SPL boundaries	SENT task
Soft	SPL $<$ 79 dB	2348
Normal	79 dB $\leq$ SPL $<$ 86 dB	2525
Loud	86 dB $\leq$ SPL $<$ 93 dB	2814
Very loud	SPL $\geq$ 93 dB	2313
Total		10 000

4.1. Labels

For the classification task, we train the classifier using supervised learning based on two different labelling approaches. The first approach is based on the subjective '*intensity category label*' explained in Section 2.3 (i.e. each sample is labelled according to the target intensity category adopted by the speaker in the recordings). The second approach corresponds to first dividing the speech samples objectively into four intensity categories based on the SPL (L_{meanZF}) of each sample, and then labelling each sample with the corresponding category. The boundaries were set to such SPL values that yielded approximately the same number of instances in each of the four category. The SPL boundaries used in the SPL-based intensity category label are described in Table 2. As shown by the second column of the table, a speech signal was labelled as 'soft' when its SPL $<$ 79 dB, as 'normal' when its SPL was 79–86 dB, as 'loud' when its SPL was 86–93 dB, and as 'very loud' when its SPL $>$ 93 dB.

4.2. Feature extraction

Three widely used frequency domain representations (spectrograms, mel-spectrograms and MFCCs) were used as features, and all of them were extracted from the normalised speech signals. The speech signals were segmented into frames of 25 ms using a Hamming window and a 5 ms overlap. Spectrograms were generated using a 1024-point FFT, yielding a 513-D vector, whereas mel-spectrograms were computed using the same FFT length and 128 mel filters, resulting in a 128-D vector. For MFCCs, a 39-D vector, which comprised delta and delta-delta coefficients, was computed. To obtain feature vectors for each sentence, the three features were aggregated separately over all the frames, and two statistics (the mean and standard deviation) were computed. Thus, the resulting feature vectors for spectrogram, mel-spectrogram, and MFCC were 1026-D, 256-D, and 78-D, respectively.

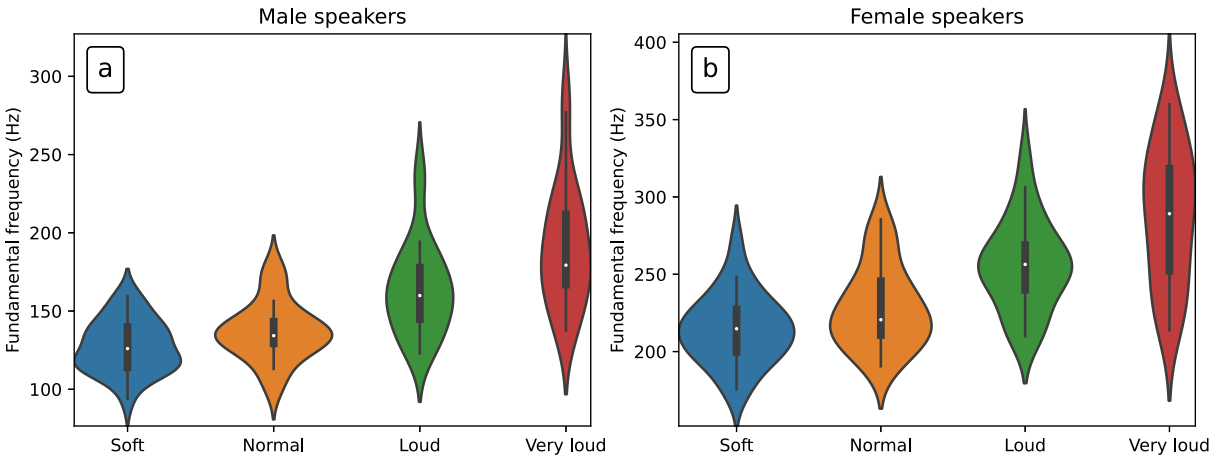

Fig. 4. Distribution of F0 for male (a) and female (b) speakers in the four intensity categories used in the recording of AVID.

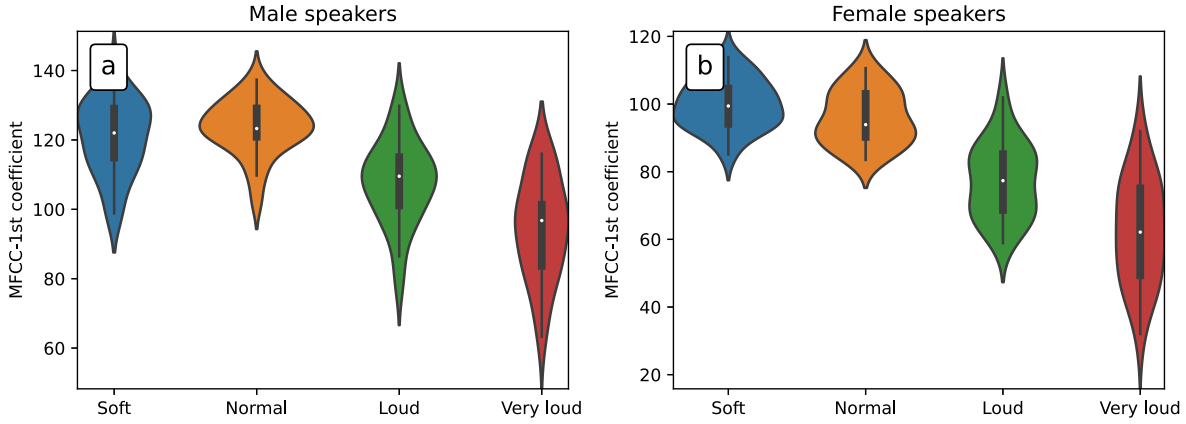

Fig. 5. Distribution of MFCC-1 (the measure of spectral tilt) for male (a) and female (b) speakers in the four intensity categories used in the recording of AVID.

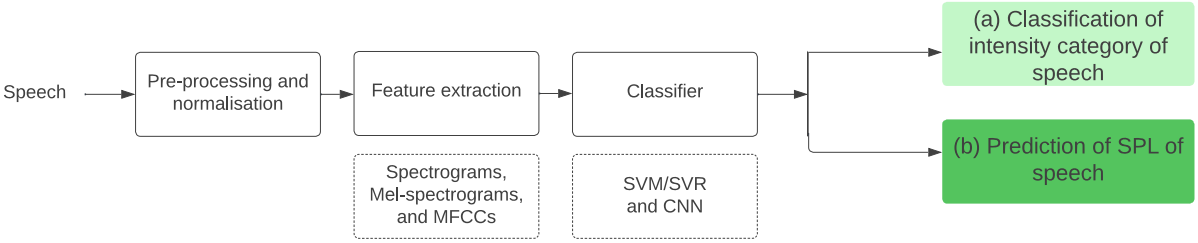

Fig. 6. Block diagram of the ML tasks studied: (a) the multi-class classification task, and (b) the regression task.

4.3. Models

For both ML tasks, we employed SVM/Support vector regressor (SVR) and CNN as ML-based models. SVM is a commonly used supervised ML algorithm for classification and regression tasks, while CNN is a widely used neural network model for classification and identification tasks. The key difference between SVM and CNN lies in the fact that SVM-based classification involves two distinct stages (feature extraction and classification), whereas in CNN, these two stages are combined into a single network, which implicitly carries out both steps. However, input hand-crafted features can also be used in CNN.

To create the training, validation and testing sets, the nested cross-validation (CV) method was used with the number of inner and outer loops set to 5 (Pedregosa et al., 2011). The GroupKFold approach was employed to divide the inner and outer loops, ensuring that the data from the same speaker was not used simultaneously in the training, validation and testing sets.

SVM was implemented using the Scikit-learn library, which uses the 'one-versus-one' approach for multi-class classification (Pedregosa et al., 2011). To optimise the hyperparameters of the SVM, GridSearchCV was used to consider a subset of three kernel types, C and gamma values. The hyperparameters used for both ML models are presented in Table 3. As there were numerous best-fitted hyperparameters for each inner loop and each setup, the resulting optimal parameter values are not reported in this article.

CNN was implemented using the Tensorflow library (Abadi et al., 2015). The Adaptive Moment Estimation (ADAM) optimiser was utilised with a learning rate of 0.001, and the batch size was set to 32. The number of training epochs was 20, and early stopping was employed to prevent overfitting. The cross-entropy function was used as the loss

Table 3
Hyperparameters used for the ML models.

Model	Hyperparameters
SVM/SVR	c : {0.1, 1, 10} gamma:{0.1, 1, 10} kernel:{'rbf', 'poly', 'sigmoid'}
CNN	conv1: filters = 32, kernel size = 3, strides = 2, activation = ReLU conv2: filters = 64, kernel size = 3, strides = 2, activation = ReLU maxpool1: pool size = 2 conv3: filters = 96, kernel size = 3, strides = 2, activation = ReLU conv4: filters = 128, kernel size = 3, strides = 2, activation = ReLU maxpool2: pool size = 2 flatten layer: - dense layer1: units = 64, activation = ReLU dropout: rate = 0.4 dense layer2: units = 4, activation = softmax

function for classification task, while mean squared error was used for regression task.

4.4. Evaluation metric

Performance of the multi-class classification systems was evaluated using accuracy as the evaluation metric and using confusion matrices to visualise misclassifications. For the regression task, the coefficient of determination (R^2), mean absolute error (MAE) and root mean square error (RMSE) were used as evaluation metrics. For an efficient regression model, R^2 should be high and MAE and RMSE should be low. In addition to these three metrics, the results of the regression experiments were assessed visually by depicting the residual plots, Q-Q plots and true *vs*. predicted plots for the testing data. The evaluation metrics were calculated for each outer loop, and then the mean and standard deviation were determined across all the loops. The evaluation metrics employed in this study are standard measures commonly utilised for assessing both classification (Géron, 2022) and regression tasks (Chicco et al., 2021).

Formulas for computing the evaluation metrics are outlined below:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \times 100\%. \quad (1)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \quad (2)$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \quad (3)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}. \quad (4)$$

where n is the number of test sentences, $\hat{y}_i$ is the predicted SPL of the ith test sentence, and y_i is the actual SPL of the ith test sentence. The mean SPL, $\bar{y}$, is given by

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

5. Results

5.1. Classification of intensity category of speech

The results of the intensity category classification experiments are reported in Table 4. The performances are presented by label (intensity category label *vs*. SPL-based intensity category label), by feature (spectrogram *vs*. mel-spectrogram *vs*. MFCCs), and by classifier (SVM *vs*. CNN). The results show that a better classification accuracy was obtained when using the objective SPL-based intensity category labels than when using the subjective intensity category label. A comparison between the features shows that, in most of the cases, the spectrogram feature performed better than the mel-spectrogram and MFCCs. Between the classifiers, there is not much difference in the performance between the labelling approaches. The best accuracy was obtained by using the SPL-based intensity category labelling, spectrogram feature and the CNN classifier, for which an accuracy of 82.0% was obtained, which is more than three times higher than the chance level (of 25%) (see Table 4).

Table 4
Mean and standard deviation of classification accuracy (in %). The values are shown separately for the two labelling approaches (intensity category label and SPL-based intensity category label), three feature sets (spectrogram, mel-spectrogram and MFCCs), and two classifiers (SVM and CNN).

Labelling	Features	SVM	CNN
Intensity category label	spectrogram	65.0 ± 2.0	64.0 ± 2.0
	mel-spectrogram	64.0 ± 1.0	63.0 ± 2.0
	MFCCs	63.0 ± 3.0	60.0 ± 3.0
SPL-based intensity category label	spectrogram	**82.0 ± 3.0**	81.0 ± 1.0
	mel-spectrogram	73.0 ± 4.0	70.0 ± 4.0
	MFCCs	65.0 ± 7.0	64.0 ± 6.0

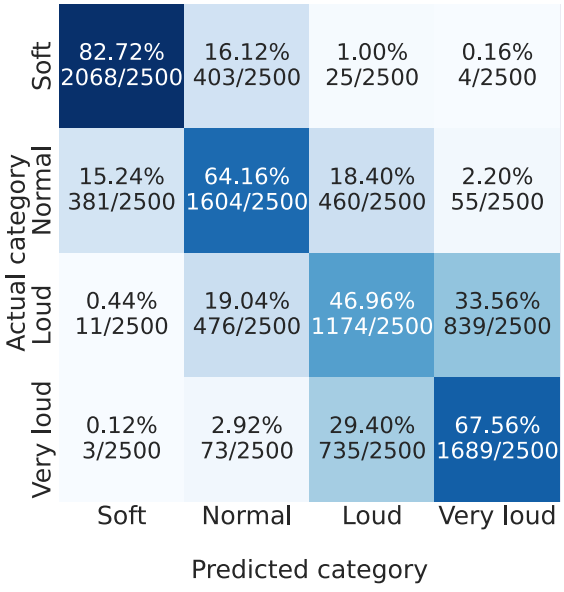

Fig. 7. Confusion matrix for the SVM classifier using the spectrogram feature based on the intensity category label. The rows are the actual intensity category, and the columns are the predicted intensity category.

The confusion matrices for the best-performing systems are depicted in Figs. 7 and 8 based on the (subjective) intensity category labelling and based on the (objective) SPL-based intensity category labelling, respectively. As can be seen from Fig. 7, the machine is able to classify the soft category best, yet there are misclassifications where soft samples are classified as normal (the neighbouring class of soft speech in Fig. 7) but clearly fewer misclassifications where soft samples are classified as loud or very loud. Using the same intensity category labelling, the machine makes the second best result for the other extreme class (very loud), yet there are more misclassifications where very loud samples have been classified into the neighbouring class (loud) compared to the number of misclassifications where soft samples have been classified into its neighbouring category (normal). In particular, correct classification of loud samples is clearly challenging for the machine when the intensity category labelling is used as indicated by the low relative number of correct classifications for the loud class (46.96%) in Fig. 7. By comparing confusion matrices of Figs. 7 and 8, it can be seen that the classification result is better for the SPL-based intensity category labelling (Fig. 8) compared to the intensity category labelling (Fig. 7)

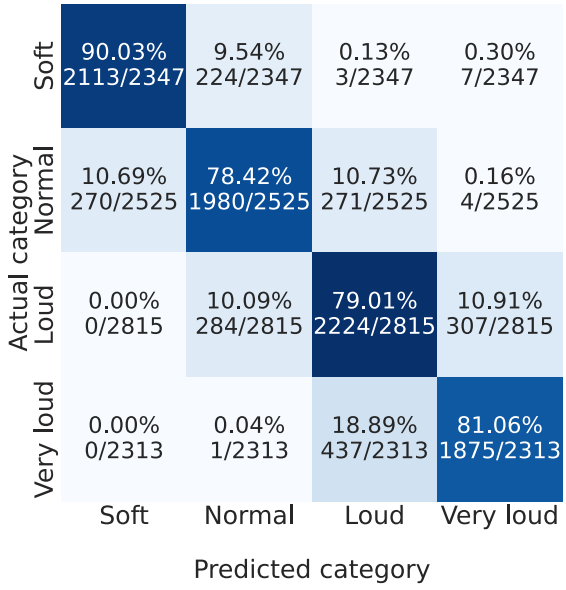

Fig. 8. Confusion matrix for the SVM classifier using the spectrogram feature based on the SPL-based intensity category label. The rows are the actual intensity category, and the columns are the predicted intensity category.

as already shown by the accuracy values reported in Table 4. From the confusion matrix shown in Fig. 8, it can be seen that the machine again makes the best result in the classification of soft samples and the second best result in the classification of very loud speech. Compared to Fig. 7, the largest improvement is achieved in the classification of normal and particularly loud speech, for which misclassifications of the correct category into the neighbouring class are clearly smaller. The differences in the observations made above from Figs. 7 and 8 can be explained by the differences in the principles of the two labelling approaches. As an example, it is namely possible that in producing loud speech (the class with the largest number of misclassifications in Fig. 7) a speaker might have used softer voice compared to another speaker's normal speech. It is difficult for the machine to make correct classifications when the labelling is based on pre-defined four intensity categories but the speaker is able to use his/her habitual regulation of vocal intensity. However, when the labelling is based on the categories defined by the objective SPL boundaries (as explained in Section 2.3), the mapping form the spectral features extracted by the machine to the SPL of the produced sound is more straightforward, which helps the machine to distinguish speech better into the four intensity categories.

5.2. Prediction of SPL of speech

The results of the regression task are reported in Table 5. The performances are presented by feature (spectrogram, mel-spectrogram and MFCCs) and by classifier (SVR and CNN). For SVR, the best metrics were achieved using the spectrogram feature, which yielded an R^2 of 92%, an MAE of 1.78 and an RMSE of 2.36. For CNN, the best metrics were obtained using the spectrogram feature that gave an R^2 of 91%, an MAE of 1.83 and an RMSE of 2.44. Comparing between the features, the spectrogram performed better than mel-spectrogram and MFCCs. Between the classifiers, SVR performed clearly better than CNN.

Figs. 9 and 10 show the residual, Q-Q and true *vs*. predicted SPL plots for the best SVR and CNN models (i.e. based on the spectrogram feature). The residual plot shows the residuals on the vertical axis and the predicted SPL values on the horizontal axis, providing information of the model's error across various target regions and allowing us to identify any presence of heteroskedasticity. The red horizontal line indicates no error, and any points above or below the line represent the magnitude of the error. Another method used to verify if residuals follow a normal distribution is the Q-Q plot. If the residuals follow a normal distribution, the observed quantiles plotted against the theoretical quantiles of a normal distribution should result in a straight line. The true *vs*. predicted SPL plots show how well the models predict SPL. A closely clustered plot around the line of perfect prediction (red line) indicates accurate predictions, while a scattered plot indicates inaccurate predictions.

From the residual plots shown by the leftmost panels of Figs. 9 and 10, it can be inferred that there are more positive residual values than negative ones. This indicates that the predicted SPL value was typically greater than the actual SPL value, implying that the ML models overestimated the SPL of speech. From the Q-Q plots shown in the middle panels of Figs. 9 and 10, it can be observed that most data points follow a roughly straight line, indicating a distribution that approximates normality. This suggests that, on average, the predicted SPL values are close to the actual SPL values. However, the presence of outliers in the plot indicates that there are specific instances where the model's predictions deviate from this average behaviour. These outliers are likely a result of the overestimation bias observed in the residual plots. As shown by the true *vs*. predicted plots depicted in the rightmost panels of Figs. 9 and 10, the data points are fairly closely clustered around the red line (i.e. the line corresponding to perfect prediction). It is worth observing that the true SPL values in these figures cover a wide dynamic range of about 60 – 105 dB, which means that the models have succeeded well in the prediction of SPL from speech signals of greatly different intensity characteristics. In addition, we would like to point out that the average absolute error between the red line and the predicted data samples in the rightmost panels of Figs. 9 and 10 correspond to the MAE values reported in Table 5 for the spectrogram feature. From the table, we can observe that that the mean MAE for the two classifiers were 1.78 and 1.83 dB. In order to assess the magnitude of these errors, we compared these current results with observations made in speech intensity measurements between healthy and pathological speech. In Fox and Ramig (1997), for example, it was reported that SPL of speech in speakers with Parkinson's disease was 2.0–4.0 lower than in speech of healthy talkers. Taken together, the results of our preliminary regression experiments indicate that the prediction of SPL from speech signals expressed on an arbitrary amplitude scale yields an average error that is smaller than the SPL difference that has been reported between healthy and pathological speech. This observation suggests that the proposed ML-based methodology could potentially be used in speech-based biomarking technology to predict SPL, a key parameter that is affected in pathological voice, from recordings that have been collected without using an SPL meter or without recording an SPL calibration tone.

6. Discussion

In recent years, many speech databases have been collected in speech technology areas such as ASR, TTS and speaker verification. However, in most popular databases, only the speech signal has been recorded without a calibration tone and without using an SLM. After the recording, speech has been stored as a digital signal whose amplitude values are expressed on an arbitrary amplitude scale that typically varies from 0.0 to 1.0. For a speech signal that has been recorded and stored this way, it is not possible to easily distinguish whether the stored sample was produced, for example, softly or loudly because the waveform has no appropriate amplitude domain information. In addition, it is not possible to know afterwards what SPL of the sample was in the recordings.

In this article, we studied the use of ML in the estimation of intensity characteristics of speech signals that are expressed as described above using arbitrary amplitude scales without calibration information. The topic is justified because the intensity regulation mechanism of speech affects not only the level of sound amplitude (as in audio equipment)

Table 5
Mean and standard deviation of R^2 (in %), MAE (in dB) and RMSE (in dB). The values are shown separately for the three feature sets (spectrogram, mel-spectrogram and MFCCs) and two classifiers (SVR and CNN).

Features	SVR			CNN		
	R^2	MAE	RMSE	R^2	MAE	RMSE
spectrogram	92.0 ± 2.0	1.78 ± 1.6	2.36 ± 2.2	91.0 ± 2.0	1.83 ± 0.16	2.44 ± 0.26
mel-spectrogram	85.0 ± 5.0	2.51 ± 3.6	3.19 ± 4.5	83.0 ± 6.0	2.66 ± 0.36	3.39 ± 0.5
MFCCs	80.0 ± 7.0	3.10 ± 0.4	3.72 ± 0.5	75.0 ± 11.0	3.18 ± 0.66	4.03 ± 0.75

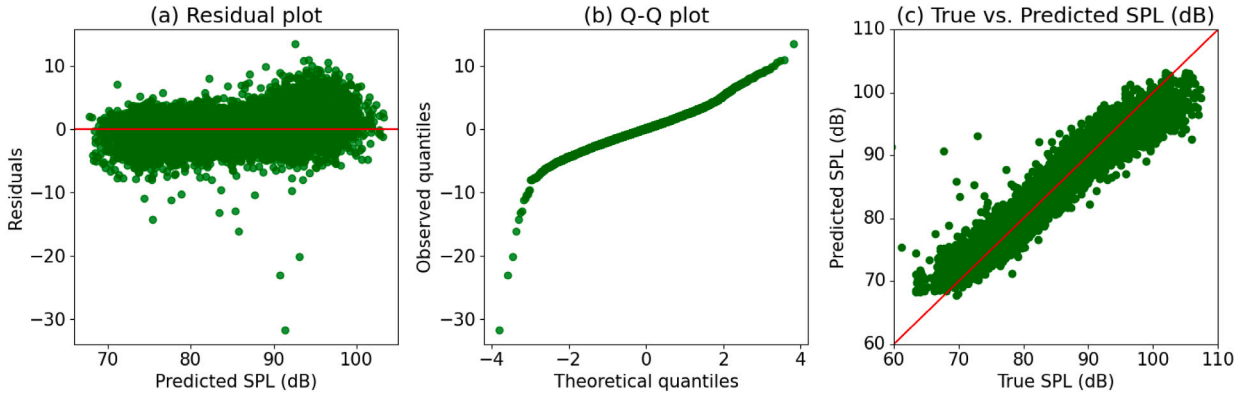

Fig. 9. Residual, Q-Q and true *vs.* predicted SPL plots for the SVR model using the spectrogram feature.

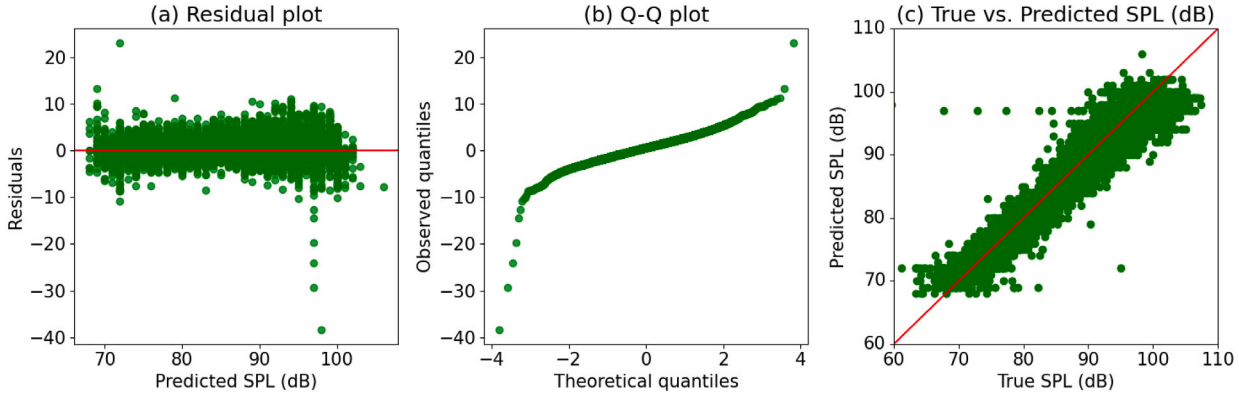

Fig. 10. Residual, Q-Q and true *vs.* predicted SPL plots for the CNN model using the spectrogram feature.

but also several other acoustical features of the signal. In order to study the topic, a new open repository called the AVID database was recorded. The database includes speech and EGG signals produced in four intensity categories (soft, normal, loud and very loud) by 25 female and 25 male speakers. Two speaking tasks (sentence reading and paragraph reading) were used to elicit the data. The data was labelled using 19 labels: the intensity category label (referring to the intensity class used by the speaker) and 18 SPL labels (referring to the sample's measured SPL including a variety of standard settings). Statistical analyses conducted on three key acoustic speech features (SPL, F0 and MFCC-1) indicated that the speakers changed their voice production manner considerably between the four target intensity categories and produced signals of greatly different acoustical characteristics.

The main goal of the study was to demonstrate how the AVID database can be used in investigating ML-based prediction of intensity characteristics of speech. Using the collected speech data, one 4-class classification task (automatic classification of intensity category) and one regression task (prediction of SPL) were studied. Both tasks were based on the scenario described above, that is, speech training and test data were represented by sentence-long speech signals whose amplitude values were deliberately warped by normalising each sample to have its waveform maximum amplitude equal to 1.0. Both ML tasks used supervised learning in which the ground truth (i.e. intensity category and SPL) was taken from the corresponding labels of AVID. In addition, both tasks used popular spectral features (spectrogram, mel-spectrogram and MFCCs) and popular ML models (SVM and CNN). By using these classifier architectures, we aimed to find out the performance achievable with state-of-the-art classifiers in the two tasks so that the results could be used as reference performance by other researchers studying the same ML tasks with the AVID data. The results of the automatic classification experiments indicated that the best accuracy (of 82%) was obtained using the spectrogram feature with the SVM classifier when the intensity category was labelled with the objective SPL-based intensity category label. Interestingly, the best accuracy obtained in the same classification task based on the subjective intensity category labels was lower (by 10% in absolute accuracy). This result indicates that the prediction of the speaker's habitual intensity category is a more difficult problem for a machine than the prediction of the intensity category, which is defined objectively based on the SPL of the produced speech signal.

For predicting SPL, the results reported in Table 5 show that the best system, which was based on the spectrogram feature and the SVR model, yielded an MAE of 1.78 dB. A comparison between this result and a previous study investigating parkinsonian and healthy speech (Fox and Ramig, 1997) shows that the mean SPL error of the best ML-based regressor of the current study was smaller than the SPL

difference between parkinsonian and healthy speech reported in Fox and Ramig (1997). This observation gives preliminary positive evidence for the utilisation of the studied ML-based methodologies in speech-based health technology in conditions where the key parameter of voice production, SPL, needs to be estimated from speech that is expressed on arbitrary amplitude scales. In addition to Parkinson's disease, SPL has been reported to be affected in several other pathologies (such as heart failure (Mittapalle et al., 2022), glottic carcinoma (Jotic et al., 2012) and cognitive impairment (De Stefano et al., 2021; Meilán et al., 2020)) as well as in healthy voice when speakers are subject to vocal loading (Laukkanen and Kankare, 2006; Södersten et al., 2002). Therefore, we believe that the proposed ML-based approach to estimate SPL in scenarios where neither SLM nor calibration information is available has lots of potential to be used in studying speech-based biomarking of different disorders, in speech therapy and in studying vocal loading. To our knowledge, the regression task has not been studied before. Therefore, the regression results obtained can be used as baselines in future studies, where more advanced ML models are developed in order to improve the prediction performance in the regression problem.

In the current article, the new AVID database was introduced and its usage in the ML-based study of vocal intensity was preliminary demonstrated. It is worth noting that the evaluations were conducted using the popular CV approach (see Section 4.1.3), in which both the system training and testing phases utilised samples from the same database (i.e. the AVID repository). In other words, there was no mismatch between the system training and testing in terms of issues such as environmental noise (i.e. all the samples were recorded in the same quiet listening room), language (i.e. all the samples were produced in English) or state of health (i.e. all the samples were produced by healthy young speakers). Therefore, a potential limitation of the current study is that the ML experiments were conducted in non-mismatched conditions and the results obtained do not provide direct evidence whether the methodology can be generalised to real-word scenarios where the test samples might be noisy or pathological, or they may represent different languages. Further research is therefore required to study how much the classification and regression performance reported in this study is affected when, for example, systems trained with the AVID samples are tested with speech data from other datasets. Moreover, as described in Section 4.2.1, the ML experiments reported in this article were conducted using three popular acoustic features (spectrograms, mel-spectrograms and MFCCs) that are all frequency-domain approaches. These features were selected because they have been widely used in multi-class classification and regression ML tasks in different areas of speech research. Moreover, since the main purpose of the present article was to introduce the new AVID database and to advertise the use of the data in ML-based research of vocal intensity, the authors did not consider it necessary to include other types of features in the current study. However, after the submission of the present article, the authors have continued their ML-based studies of vocal intensity by publishing a recent conference article (Kodali et al., 2023a). In this conference paper, the authors study the same intensity category classification task as in the current article by comparing the three spectral features with embeddings from two pre-trained models (Wav2vec2 (Baevski et al., 2020) and Whisper (Radford et al., 2022)) and using SVM as classifier. The results of these latest experiments show that the pre-trained model embeddings outperform the spectral features by providing an improvement of about 7% (absolute) in accuracy.

Compared to the popular databases that are used in major areas of speech technology, the proposed AVID database is clearly smaller. As described in Section 2.1, AVID includes 16 h of raw speech produced by 50 speakers, whereas the amount of speech and the number of speakers in, for example, TIMIT (Anon, 1993), LibriSpeech (Anon, 2022), and NIST Speaker Recognition Evaluation Test Set (Greenberg et al., 2019) are considerably larger (630 speakers and 5 h of speech in TIMIT; 2500 speakers and 1000 h of speech in LibriSpeech; 220 speakers and 340 h of speech in NIST). The main significance of AVID is the availability of intensity information (i.e. SPL values, intensity category) that is lacking in most popular databases. Compared to a few speech databases that have been used in similar studies on vocal intensity, the size of AVID data is, however, much larger. The dataset used in Zhang and Hansen (2007) and Zelinka et al. (2012), for example, included 12 and 13 speakers, respectively, and all the speakers in both datasets were male.

7. Conclusions

This study introduced the AVID database, which includes calibrated speech recordings produced by 50 speakers in four intensity categories. As the main contribution, the study demonstrated that the AVID data can be used to train ML models to estimate both the intensity category and the SPL of speech in scenarios where the original level information of speech is not available in the signal waveform. Therefore, the proposed ML approaches can in principle be used to estimate the intensity characteristics of speech samples of existing databases that have been collected without recording the calibration tone. In addition, the proposed ML methodologies can be used in the estimation of speech intensity characteristics of new recordings that are conducted in circumstances where the use of an SLM or recording of a calibration tone is not possible (e.g. in recording speech of patients with a phone or laptop in real-life conditions outside clinics). In particular, the authors would like to point out that the best regression network gave an MAE that was less than 2 dB in the SPL prediction task despite the original level information of speech was deliberately removed. The authors consider this achievement particularly important and it may have far-reaching implications to the speech-based biomarking study of health because it suggests that the studied ML technology could predict SPL of speech with a precision that enables distinguishing healthy and pathological voices despite their true SPL levels have not been measured with an SLM. The AVID database is publicly available at Kodali et al. (2024). The authors welcome researchers interested in intensity regulation of speech to utilise the database. In particular, we encourage researchers interested in classification and regression problems to utilise AVID in the development of new deep-learning methods for the study of vocal intensity.

CRediT authorship contribution statement

Paavo Alku: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition. **Manila Kodali:** Data curation, Methodology, Visualization, Writing – original draft, Writing – review & editing, Software. **Laura Laaksonen:** Validation, Visualization, Funding acquisition. **Sudarsana Reddy Kadiri:** Validation, Supervision, Visualization, Writing – review & editing.

Declaration of competing interest

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.

Data availability

Data available at https://zenodo.org/doi/10.5281/zenodo.7948299.

Acknowledgements

This study was funded by the Academy of Finland (project no. 330139) and Huawei Finland. Aalto ScienceIT provided the computational resources.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org, URL https://www.tensorflow.org/.

Alku, P., Airas, M., Björkner, E., Sundberg, J., 2006. An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity. J. Acoust. Soc. Am. 120, 1052–1062.

Alku, P., Murtola, T., Malinen, J., Kuortti, J., Story, B., Airaksinen, M., Salmi, M., Vilkman, E., Geneid, A., 2019. OPENGLOT–An open environment for the evaluation of glottal inverse filtering. Speech Commun. 107, 38–47.

Alku, P., Vintturi, J., Vilkman, E., 2002. Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation. Speech Commun. 38 (3–4), 321–334.

Anon, 1993. TIMIT acoustic-phonetic continuous speech corpus. https://catalog.ldc.upenn.edu/LDC93S1. (Online; Accessed: 28 May 2022).

Anon, 2008. The call of the wild by Jack London. https://www.gutenberg.org/ebooks/215. (Accessed: 30 June 2021).

Anon, 2021a. Amprobe sound meter calibrator home page. https://www.amprobe.com/product/sm-cal-1/. (Accessed: 30 October 2021).

Anon, 2021b. RME Babyface sound card home page.. https://babyface.rme-audio.de/. (Accessed: 5 September 2021).

Anon, 2021c. Weather forecasting excerpt. https://bit.ly/3iDF3K6. (Accessed: 30 June 2021).

Anon, 2022. LibriSpeech ASR corpus. https://www.openslr.org/12. (Online; Accessed: 28 May 2022).

Baevski, A., Zhou, Y., Mohamed, A., Auli, M., 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. Adv. Neural Inf. Process. Syst. 33, 12 449–12 460.

Baghel, S., Prasanna, S.R.M., Guha, P., 2020. Exploration of excitation source information for shouted and normal speech classification. J. Acoust. Soc. Am. 147, 1250–1261.

Barras, B., 2012. SoX: Sound eXchange. Tech. rep..

Chicco, D., Warrens, M.J., Jurman, G., 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput. Sci. 7, e623.

Clark, J.P., Adams, S.G., Dykstra, A.D., Moodie, S., Jog, M., 2014. Loudness perception and speech intensity control in parkinson's disease. J. Commun. Disorders 51, 1–12.

Coleman, R.F., Mabis, J.H., Hinson, J.K., 1977. Fundamental frequency-sound pressure level profiles of adult male and female voices. J. Speech Hear. Res. 20, 197–204.

De Stefano, A., Di Giovanni, P., Kulamarva, G., Di Fonzo, F., Massaro, T., Contini, A., Dispenza, F., Cazzato, C., 2021. Changes in speech range profile are associated with cognitive impairment. Dementia Neurocogn. Disord. 20 (4), 89.

DeKeyser, K., Santens, P., Bockstael, A., Botteldooren, D., Talsma, D., DeVos, S., Cauwenberghe, M.V., Verheugen, F., Corthals, P., DeLetter, M., 2016. The relationship between speech production and speech perception deficits in parkinson's disease. J. Speech Lang. Hearing Res. 59 (5), 915–931.

Fox, C.M., Ramig, L.O., 1997. Vocal sound pressure level and self-perception of speech and voice in men and women with idiopathic parkinson disease. Am. J. Speech-Lang. Pathol. 6, 85–94.

Garofolo, J.S., 1993. TIMIT acoustic phonetic continuous speech corpus. In: Linguistic Data Consortium, 1993.

Géron, A., 2022. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, Inc.

Greenberg, C.S., Masonb, L.P., Sadjadia, S.O., Reynolds, D.A., 2020. Two decades of speaker recognition evaluation at the national institute of standards and technology. Comput. Speech Lang. 60, 101032.

Greenberg, C., Sadjadi, O., Kheyrkhah, T., Jones, K., Walker, K., Strassel, S., Graff, D., 2019. 2016 NIST speaker recognition evaluation test set. URL-http://dx.doi.org/11272.1/AB2/WJ2G5L.

Herbst, C.T., 2020. Electroglottography–an update. J. Voice 34 (4), 503–526.

Hodge, S., Colton, R., Kelley, R., 2001. Vocal intensity characteristics in normal and elderly speakers. J. Voice 7, 503–511.

Holmberg, E., Hillman, R., Perkell, J., 1988. Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice. J. Acoust. Soc. Am. 84, 511–529.

Huber, J.E., Stathopoulos, E.T., Curione, G.M., Ash, T.A., Johnson, K., 1999. Formants of children, women, and men: The effects of vocal intensity variation. J. Acoust. Soc. Am. 106, 1532–1542.

Jotic, A., Stankovic, P., Jesic, S., Milovanovic, J., Stojanovic, M., Djukic, V., 2012. Voice quality after treatment of early glottic carcinoma. J. Voice 26 (3), 381–389.

Kibria, S., Samin, A.M., Kobir, M.H., Rahman, M.S., Selim, M.R., Iqbal, M.Z., 2022. Bangladeshi bangla speech corpus for automatic speech recognition research. Speech Commun. 136, 84–97.

Kodali, M., Alku, P., Kadiri, S.R., 2024. AVID: Aalto vocal intensity database. URLhttps://dx.doi.org/10.5281/zenodo.7948300.

Kodali, M., Kadiri, S., Alku, P., 2023a. Classification of vocal intensity category from speech using the wav2vec2 and whisper embeddings. In: Proc. Interspeech. pp. 4134–4138.

Kodali, M., Kadiri, S.R., Laaksonen, L., Alku, P., 2023b. Automatic classification of vocal intensity category from speech. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 1–5.

Laffitte, P., Sodoyer, D., Tatkeu, C., Girin, L., 2016. Deep neural networks for automatic detection of screams and shouted speech in subway trains. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing.

Laukkanen, A.-M., Kankare, E., 2006. Vocal loading-related changes in male teachers' voices investigated before and after a working day. Folia Phoniatr. Logop. 58 (4), 229–239.

Liénard, J.-S., 2019. Quantifying vocal effort from the shape of the one-third octave long-term-average spectrum of speech. J. Acoust. Soc. Am. 146, EL369–EL375. http://dx.doi.org/10.1121/1.5129677.

Lienard, J.-S., Benedetto, M.-G.D., 1999. Effect of vocal effort on spectral properties of vowels. J. Acoust. Soc. Am. 106, 411–422.

Meenakshi, G.N., Ghosh, P.K., 2015. Robust whisper activity detection using long-term log energy variation of sub-band signal. IEEE Signal Process. Lett. 22, 1859–1863.

Meilán, J.J., Martínez-Sánchez, F., Martínez-Nicolás, I., Llorente, T.E., Carro, J., et al., 2020. Changes in the rhythm of speech difference between people with nondegenerative mild cognitive impairment and with preclinical dementia. Behav. Neurol. 2020.

Mittapalle, K.R., Pohjalainen, H., Helkkula, P., Kaitue, K., Minkkinen, M., Tolppanen, H., Nieminen, T., Alku, P., 2022. Glottal flow characteristics in vowels produced by speakers with heart failure. Speech Commun. 137, 35–43.

Nash, A., 2014. An electronic database of speech sound levels. In: Proc. Inter-Noise. pp. 4296–4302.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pohjalainen, J., Raitio, T., Yrttiaho, S., Alku, P., 2013. Detection of shouted speech in noise: Human and machine. J. Acoust. Soc. Am. 133, 2377–2389.

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., 2022. Robust speech recognition via large-scale weak supervision. arXiv:2212.04356.

Rec, I., 1997. BS. 1116-1: Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. Int. Telecomm. Union, Geneva Std.

Sarria-Paja, M., Falk, T.H., 2013. Whispered speech detection in noise using auditory-inspired modulation spectrum features. IEEE Signal Process. Lett. 20, 783–786.

Schulman, R., 1989. Articulatory dynamics of loud and normal speech. J. Acoust. Soc. Am. 85, 295–312.

Sharma, B., Gao, X., Vijayan, K., Tian, X., Li, H., 2021. NHSS: A speech and singing parallel database. Speech Commun. 133, 9–22.

Södersten, M., Granqvist, S., Hammarberg, B., Szabo, A., 2002. Vocal behavior and vocal loading factors for preschool teachers at work studied with binaural DAT recordings. J. Voice 16 (3), 356–372.

Švec, J.G., Granqvist, S., 2018. Tutorial and guidelines on measurement of sound pressure level in voice and speech. J. Speech Lang. Hearing Res. 61, 441–461.

Švec, J.G., Titze, I.R., Popolo, P.S., 2005. Estimation of sound pressure levels of voiced speech from skin vibration of the neck. J. Acoust. Soc. Am. 117 (3), 1386–1394.

Titze, I., 1994. Principles of Voice Production. Prentice-Hall, NJ.

Titze, I., Sundberg, J., 1992. Vocal intensity in speakers and singers. J. Acoust. Soc. Am. 91, 2936–2946.

Traunmüller, H., Eriksson, A., 2000. Acoustic effects of variation in vocal effort by men, women, and children. J. Acoust. Soc. Am. 107 (6), 3438–3451.

Wielgat, R., Jędryka, R., Lorenc, A., Mik, Ł., Król, D., 2021. POLEMAD–a database for the multimodal analysis of polish pronunciation. Speech Commun. 127, 29–42.

Zelinka, P., Sigmund, M., Schimmel, J., 2012. Impact of vocal effort variability on automatic speech recognition. Speech Commun. 54 (6), 732–742.

Zhang, C., Hansen, J.H.L., 2007. Analysis and classification of speech mode: Whispered through shouted. In: Eighth Annual Conference of the International Speech Communication Association. pp. 2396–2399.

Zhang, C., Hansen, J.H.L., 2011. Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing. IEEE Trans. Audio Speech Lang. Process. 19, 883–894.

Zhou, K., Sisman, B., Liu, R., Li, H., 2022. Emotional voice conversion: Theory, databases and ESD. Speech Commun. 137, 1–18.