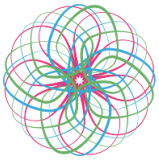
Universiteit Leiden



CHDR

Master Computer Science

Comparison of Speaker Verification Models Using Smartphone Recordings Collected in Healthy Human Subjects

Name:	Zsolt Hubai
Student ID:	s2599465
Date:	27 / 07 / 2021
1st supervisor:	Erwin M. Bakker (LIACS)
2nd supervisors:	Ahnjili ZhuParris & Robert J. Doll (CHDR)

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)

Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

For the widespread use of speaker verification models, evaluation is important on out-of-domain datasets, meaning that the source of the utterances for the verification phase is significantly different than the training dataset. The technical limitations of smartphone microphones and many other environmental obstacles (e.g., background noise) creates a demand of robust smartphone-based speaker verification systems. Using an Android application, human speech is collected by researchers to induce changes in vocal biomarkers. The use of speaker verification could support this application by automatically detecting a speaker of interest from the recordings to deepen the scope of analysis and circumvent handling potentially sensitive data. The goal of this study is to find optimal machine learning model for a future deployment of the speaker verification system on smartphone recordings We compare three state-of-the-art speaker verification systems and evaluate the performance on a new dataset recorded with smartphones. We verify 16 female and 16 male speakers in different environmental conditions, identify the optimal sampling rate and compare varying enrolment and test input lengths. The results show that the models underperform in scenarios that are considered challenging but show promise in less noisy conditions. We observed the effects of different input lengths, gender, and recording conditions but not the speech-content on the performance. From the investigated models, overall, the best performing model is a ResNet34-based meta-learner system which achieved the lowest equal error rate. However, a computational efficiency analysis showed a SincNet X-vector model yielding fastest results. Additional necessary research is defined in experimenting in less standardised conditions with increased session variability for each speaker. We found the ResNet34 based Meta-Learning model trained specifically to process both short and long utterances to be the optimal model for smartphone recordings in a clinical trial setting.

Contents

1.INTRODUCTION

Smartphone devices are becoming a popular tool in many medical industries where mass data collection is carried out on human subjects. Clinical trials are one of those areas where smartphone devices are used to collect information on patients remotely, unobtrusively, and effectively. The built-in microphone of modern smartphone devices can be used to collect human speech data from the participants to measure intervention induced changes in vocal biomarkers such as tone or pitch [70]. Centre for Human Drug Research (CHDR) is a Netherlands- based institute specializing in early-phase clinical drug research. The CHDR Trial@home® program enables clinical trials to be conducted remotely. The technology supporting Trial@home® is the CHDR MORE® application that can be integrated into digital devices, including Android smartphones. The application can collect, store, transform and transfer data originating from various smartphone sensors, such as accelerometers, gyroscopes, camera, and microphone. It has been used for analysis in validated studies in areas like neurology, psychiatry, and dermatology [1]. The application has an integrated Voice Activity Detection (VAD) module that detects human speech within a certain vicinity of the smartphone. However, VADs have limited capabilities in identifying individual speakers, and therefore cannot distinguish between the speaker of interest and their peers. The scope of possible analysis is limited to only approximating the quantity of social interactions of a subject but not accurately measuring his/her engagement. Extra burden is placed on the researchers by having to manually listen to the recordings to verify and distinguish the subject from the environment. The motivation of our study is to improve the capabilities of the platform by integrating a speaker recognition model that could verify and detect the identities of the users based on their voices. Since capturing the exact spoken words is not the intention nor is it necessary to recognize changes in speech, the automatic process would circumvent the need to gather sensitive data. Speaker recognition could assist the application by verifying the exact identity of the user (trial subjects/ patients) recordings and would enable the automatic processing of the data without manual annotation, and without the invasion of the subject's privacy.

Speaker recognition (SR) is a broad research field that focuses on speaker modelling from voice features that relate to unique speaker characteristics [2]. SR is separate from automatic speech recognition (ASR) which focuses more on the content of the speech or the features related to the audio signal. Recent academic studies agree on differentiating three sub-fields for SR: Speaker Verification (SV), Speaker Identification (SI) and Speaker Diarization (SD) [3]. An overview of the hierarchical composition of SR can be seen in Figure 1. SV is the task of deciding whether two voice samples originate from the same speaker or not [4]. In other words, it must verify if the spoken words, statement, or vocal sound, called utterances belongs to our speaker of interest or to a potential intruder. The matching is based on the speaker's already known utterances [5]. SV is conducted in an open-set fashion, meaning that the number of speakers it encounters is not fixed in advance (it must be able to successfully compare previously unseen speakers). There is a growing interest in researching SV mainly due to the popularity of voice assistant applications that partly apply reliable SV systems (e.g., Apple Siri or Amazon Alexa) [6]. SI is very similar to SV, in which multiple reference utterances are compared to the voice signal in a closed-set environment [3]. The main difference is that SI is considered a classification problem while SV is a decision problem at its core. Finally, SD is the task of answering the question "who spoke when?" [7] by labelling a longer audio stream into homogeneous speaker segments [8].

SR can be text-dependent or text-independent. In the text-independent setting, no lexical constraints for the utterances to match (e.g., "OK Google") are set [5]. The speakers are allowed free-form voice input; thus, these types of systems are most common in forensic and cyber-security applications. This task is considered more challenging than the text-dependent scenario [9]. The text-independent SV approach was chosen to model and verify speakers for clinical trials due to the open-set, generalisation property and the ability of free speech input.

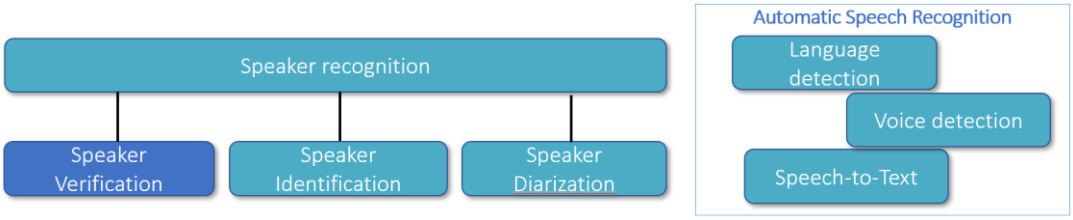


Fig. 1. Overview of automatic speaker recognition, which is part of a larger field called automatic speech recognition.

Before describing the SV process in more detail, we introduce human voice production and the subsequent digital coding of sound. The vocal tract system is responsible for generating speech, which consists of phonemes that represent the sounds of a language. The opening and closing of the vocal cords generate periodic pulses of air pressure that energizes the vocal tract tube. Different types of phonemes are the consequence of different ways the air flows and resonates through the vocal tract shapes. The frequency response of the vocal cords and tract gives phonemes a certain resonance frequency which makes up a sound [56]. It must be noted that the lengths of the vocal tract differ for males and females, which affects the sound production. The analogue sound signal reaches a capturing device, e.g., a smartphone microphone. It is converted to a digital signal by repeatedly measuring the energy level of the signal at certain points in time. These short time intervals are defined by the sampling rate, the number of samples captured per second (measured in Hz or kHz). The quantization of the sampled analogue signal is done by a method called Pulse-code modulation (PCM).

SV systems are required to represent each speaker uniquely, independent of the conditions. It must recognise the same speaker across different timeframes and environments. The model also must be discriminative: if the input samples originate from different speakers, we expect the representations to be dissimilar. These two properties are achieved by training a so-called universal background model (UBM) using a large set of diverse speaker database. UBM is a framework which allows the representation of general speaker characteristics, independent of a specific person. The background model is then used as a feature extractor during the verification process for the enrolment and test utterances. Due to UBMs being universal, they can be fitted to any speaker without retraining the model which is advantageous. Creating a large training set carefully catered for the future application is not feasible in most cases because it needs multiple samples from thousands of speakers. The basic components of building SV models are *development*, *enrolment*, and *testing*. During the *development* phase, an algorithm is trained during a classification task of known speakers to create discriminative speaker representations. The final speaker embeddings are dependent on the used method (e.g., Gaussian mixtures or Deep Neural Networks), the level of

aggregation (frame or utterance) and are directed by the loss function: cross-entropy or metric learning [53]. In essence, speaker embeddings are vectors coding voice characteristic of speakers. The background model is used as a feature extractor for the verification phase which consists of *enrolment* and *testing*.

The *enrolment* and *testing* phases are different only when applying SV in a real-life application but are technically the same process. Enrolment is the process when new, previously unseen voice samples from a speaker or multiple speakers are fed to the trained model. The resulting speaker embeddings are stored to serve as references when unknown samples are captured. This happens during testing when both embeddings of the enrolled speaker(s) and unknown/ impostor speaker(s) are extracted. The distance of these embedding vectors compares the likeliness of the original voice samples and the new voice samples. This distance metric is used as a scoring function, most commonly the cosine similarity [3]. Finally, a decision whether to accept two samples as belonging to the same speaker or not is based on a threshold value . The false rejection rate and the false acceptance rate is taken into consideration when choosing the threshold value and is always dependent on the actual application [68]. A high-level overview of the verification process is shown in Figure 3.

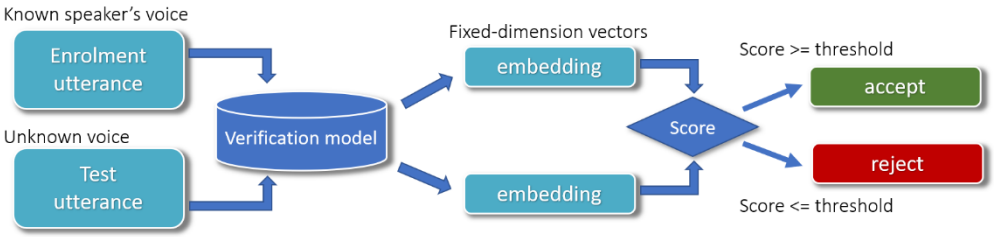


Fig. 3. High-level overview of the verification process.

The usual SV pipeline requires two different datasets, a development dataset, and a verification dataset. SV models are trained using a development dataset, a large, labelled corpus of many speakers (1000+) of many different characteristics to adequately capture intra-and inter-speaker variations. The verification dataset is an 'in-domain' dataset meaning that it is recorded in the same or very similar conditions as the future application. It is usually much smaller making it ineffective to use for training (or for fine-tuning). Usually, SV models are trained and evaluated using datasets from the same domain. Out-of-domain dataset means that a substantial mismatch exists between the development and the verification set in terms of recordings conditions, types of speakers, language. Performance of an SV model drops when feeding it data from a different domain [69]. In most cases, the literature focuses on building new feature extractor models to achieve better performance on a few popular publicly available speaker datasets extracted from voice recordings (TV/film production, audiobooks) [12]. No thorough research has been done on the performance of these models on smartphone recordings. We define smartphone recordings as being lower quality data due to the limited technological capacity of their built-in microphones and showing variable distance between the microphone and the speaker. We assume that models optimised for non-smartphone recordings would perform subpar and thorough research is necessary for this novel domain. Existing studies have only focused on the adaption of SV system to mobile devices by increasing algorithmic efficiency [13,14].

The contribution of this study is an adequate comparison of modern SV models, with respect to their performance on a smartphone-based dataset. Specifically, we investigate the ability of SV models trained on public datasets to perform SV on smartphone recordings for different scenarios. Data collection for clinical trials are not always standardised so the models should be robust to different acoustical environments, speakers, and context. We investigate the models' performances when the recording device is located relatively close-up to the speakers, far away from him/her, or is inside the pocket of said subject. We also add 'media' speakers, speech that could originate from third-party devices (TV, loudspeakers). We apply different sampling rates to the recordings to investigate the performance-storage requirement trade-off. We also analyse the models by their effectiveness, the speed with which they operate. The evaluation process is carefully constructed so that is resembles a future real-life usage. The rest of the paper is organized as follows. In Section 2, traditional methods are discussed followed by an in-depth overview of the deep learning-based models. Section 3 explains the datasets, the features, the architecture of the specific models used for this study, and the evaluation strategy. In Section 4, we show the results of all corresponding experiments. Different pre-processing scenarios are analysed (sampling rate, enrolment length, noise levels) on three different recording conditions. Further experiments are made with respect to the performance of the models of diverse types of speakers (directly speaking to the microphone vs voices played through a loudspeaker), gender and the type of speech. Finally, an analysis is made of the computational efficiency of the models with a recommendation on a future use for clinical trials and remote patient tracking applications. Sections 5 and 6 of the thesis conclude the study and discuss limitations and future work.

2. BACKGROUND AND RELATED WORK

In this section an overview of machine learning and deep learning models used in SV is provided. We also give a comprehensive outline of different deep learning models from various aspects such as architecture, objective function, input features, and training dataset. Before 2010, Gaussian Mixture Models (GMM) were considered state-of-the-art for modelling speakers [15]. A fixed number of multivariate Gaussians form the GMM and the expectation–maximization algorithm used to train the (UBM) from a large set of speakers [4]. The likelihood of an unknown speaker matching a known speaker is the difference between the speaker specific GMM and the UBM [67]. The first problem arises with these traditional approaches is intra-speaker variability attributed to the changing phonetic content and coarticulation. The second problem is channel variability due to environmental circumstances like background noise [66]. Maximum a Posteriori (MAP) adapted GMMs were used to mitigate speaker and channel variability using latent factor analysis where the means of the mixture components are stacked to create a supervector [16]. Further developments based on the GMM-UBM method were the integration of SVMs (Support Vector Machine) and Joint Factor Analysis for estimation [17,18]. Later, a new modelling technique called 'i-vector' (identity vector) was presented to represent each utterance in a low-dimensional space instead of a high-dimensional GMM mean supervector space [19]. Text-independent verification especially benefited from this property because information related to text is simply scrapped from the more concise representation of i-vectors [20]. This new feature extraction front-end technique was combined with Probabilistic Linear Discriminant Analysis (PLDA) verification back-end creating the new state-of-the-art [21]. Although these traditional methods achieved reasonable performance, encoding irrelevant information related to noise and channel remained a

problem which forced researchers to keep exploring more representative and discriminate features.

2.1 Deep neural architectures

The recent success of deep learning in ASR motivated researchers of SR to replace i-vectors with speaker embeddings extracted from hidden layers of a neural network. The obtained frame-level features are called 'd-vectors' [22]. Improving on this approach, 'x-vectors' are segment-level speaker embeddings extracted from one of the deeper layers of the neural network [23]. Many deep learning architectures have been proposed for text-independent SV, the two most popular being Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). Examples for SV models incorporating CNNs include a Siamese architecture to separate same-speaker and different-speaker pairs [24], a 3D-CNN architecture with stacked input features [9] and VGGNet [25]. The first convolutional layer is often replaced by parametrized Sinc functions which improve capturing narrow-band speaker characteristics [26]. By far the most popular CNN architecture is ResNet [27] which was successfully applied to this task as well with variations of the network [28–30]. A Long Short-Term Memory (LSTM) network, an RNN-based network, with generalized end-to-end loss has been proposed by Google [31]. Effective Time Delay Neural Networks (TDNN) were shown to model long term temporal dependencies. It captures wider context inputs to create better representations [32]. Many hybrids and alternative models exist, like Wav2Vec, which is a self-supervised framework based on the recent success of transformers [33] or Wav2Spk [34]. RawNet extracts frame-level embeddings using residual blocks with CNNs, then aggregates features into utterance level using LSTM [35]. There are two main advantages for using deep learning over traditional methods (GMMs and i-vectors). First, the complex neural network can model more discriminative speaker representations because they are able to leverage information at both frame- and utterance level. This is usually achieved by a special deep layer which is responsible for the transformation. Second, the varying speech features between males and females do not affect performance if trained properly. Thus, it is not necessary to build separate models anymore for both genders [36]. In the past, the spoken language also played a role in the effectiveness of SV systems as the performance declined for foreign speakers [10]. Today's state-of-the-are SV systems are more robust in terms of language variance [11].

2.2 Development datasets

We introduce some of the large open datasets that were used by researchers in the past to train robust feature extractors and evaluate models on various challenges. The growing interest in SV research originated partly because some newer successful effort by researchers to create such public speaker database. LibriSpeech [48] is a corpus of 1000-hour English speech extracted from 2484 speakers reading audiobooks. It is originally intended for speech recognition because of the attached transcript, but it was used for text-independent SV tasks as well [49,50]. A variation of this dataset is called VOICES for which samples were played through a loudspeaker with background noise in a room to create more real-life conditions [51]. Speakers in the wild dataset was created in uncontrolled conditions and it contains hard-annotated speech samples from 299 speakers across eight sessions on average [52]. By far the most comprehensive and diverse dataset for SV is the VoxCeleb1 and the larger set VoxCeleb2. It is an audio-visual dataset used for face recognition as well as many fields in ASR. The two sets contain more than 7000 speakers and

1,000,000 utterances from different speakers of age, nationality, ethnicity, and language. It contains both clean and noisier recordings extracted from YouTube [28].

2.3 DNN inputs

Another interesting aspect of the deep-learning modes is the type of input. Cepstrum-based features were used extensively for SV just like in any other areas of speech processing. These features can represent the envelope, the shape of the vocal tract when pronouncing certain phonemes, and thus can represent unique speech and speaker characteristics. Cepstrum-based features are the mel-frequency cepstral coefficients (MFCC) [42,43] and filter-bank energies [44,45]. When extracting these features, the speech spectrum is smoothed which is a disadvantage because narrow-band speaker characteristics (pitch and formants) are harder to capture [26]. End-to-end DNN models can replace hand-crafted features all together and encode raw audio signals [26,47].

2.4 Loss functions

In deep learning models, the loss function directs the training and the quality of representations consequently, so it is important to summarize the literature from this aspect. Since SV is a classification problem, classification-based objective functions were widely used when creating the first neural network-based models [3] [23]. The minimum cross entropy objective function combined with a Softmax output layer is called the Softmax loss (Eq. 1).

$$L_{softmax} = -\log \frac{\exp(W_y x + b_y)}{\sum_{j=1}^{c} \exp(W_j^T x + b_j)} \tag{1}$$

where W_i and b_i are the weight vector and bias term before the Softmax layer corresponding to class y, x denotes the input vector, and c denotes the number of classes. Such simplistic objective function helped minimizing the between-speaker distance but not the intra-speaker variance. Improving on this objective function, a margin was introduced between classes to learn more discriminative features [37]. One of these margin-based objective functions is the Additive angular margin loss (AAM) [38]. By discarding the bias term and only considering the directions of the columns of W in Eq. 1, the expression can be rewritten as:

$$L_{softmax} = -\log \frac{\exp(||x_i|| \cos(\theta_{y_i}, i))}{\sum_{i=1}^{c} \exp(||x_i|| \cos(\theta_j, i))} \tag{2}$$

where θ_i, i is the angle between weight and input. If W_i and x_i are normalized, then the angle is calculated by:

$$\theta = \arccos(W_j^T x) \tag{3}$$

Given a scaling factor s and the margin m we end up with the definition of the Additive angular

margin loss [39]:

$$L_{AAM} = -\log \frac{\exp(s(cos(\theta_y + m)))}{\exp(s(cos(\theta_y + m))) + \sum_{j=1, j\neq y}^{c} \exp(s(cos(\theta_y)))} \tag{4}$$

The previous two objective functions were classification-based losses: each class of the training set is known. Since SV is an open set problem, the actual classification accuracy is not the interest, rather its ability to discriminate between two speakers [40]. A set of objective functions rely only on binary annotation of pairs of training samples: whether they originate from the same speaker or not. These are called contrast-based losses and they provide alternative losses for SV. Triplet loss was introduced to pull samples closer from the same class and push samples away from the different classes. It is achieved by using three training samples: a reference utterance (u_r), a positive utterance (u_p) and a negative utterance (u_n) and the loss is calculated by:

$$\text{Triplet loss} = -\sum_{\cdot\cdot=1}^{N} max(0, S_{rp} - S_{rn} + \zeta) \tag{5}$$

where S_{rp} is the similarity score between the reference and positive utterance (same speaker) and S_{rn} is the similarity score between the reference and negative utterance (different speaker) while ζ is the margin between the positive and negative pairs [41].

3. METHODS

3.1 Data

As mentioned in Section 1, SV models utilise two types of datasets: a development set and a verification set. Traditionally, the two datasets are chosen from the same domain, e.g., a simple train/test split of a coherent dataset. In contrast, this study focuses on applying SV models to a new domain where the environmental conditions, recording device, and language are different. Next, we describe these two sets, the publicly available development dataset, and the newly created verification dataset.

3.2 Development dataset

The models compared in this study were trained on the VoxCeleb2 dataset [28]. It contains short clips of human speech extracted from interview videos on YouTube sampled at 16kHz. It consists of a total 6,112 different speakers and 1,128,246 utterances. An official train-test split is given by the creators of the dataset: 5994 speakers belong to the training set and the remaining 118 speakers belong to the test set. The average number of utterances per speaker is 185 and the average length of an utterance is 7.8 seconds. Furthermore, the dataset contains speech in multiple languages, although U.S. speakers take up most of the data (29%). The exact number of Dutch speakers is not known, but it is less than 6%. The gender ratio of the speakers is 39% to 61% for females and males respectively.

3.3 Verification dataset

We created our own dataset to evaluate models recording speakers using smartphones. The data collection pipeline was designed to resemble a realistic future use, to monitor subjects in a naturalistic environment.

Participants: Speech from 8 male and 8 female Dutch speakers have been collected with three phones recording simultaneously from different positions. The participants were volunteers recruited by CHDR and had to be fluent Dutch speakers who could read above an A2 reading level. All participants gave consent to use their voice recordings for this study. No personal information was collected from the participants other than gender, age (in years) and their voice samples. The age range of the participants was between 20 and 35. Another set of voice sample were collected from speakers we call 'media-speakers' (while the previous group is named 'subject-speakers'). Speech was extracted from YouTube videos and corresponding to 8 male and 8 female Dutch speakers in the context of public speeches, podcasts, audiobooks interviews and lectures. The identities of these speakers together with the YouTube links can be found in Appendix 9.4. The purpose of this second group of speakers were to imitate a scenario when speech is captured from a device around the patient which might be captured by the microphone.

Equipment: Three Motorola g6 smartphones running on Android operating system were used to collect the speech data. For all sessions the same three devices were used in a randomized order. The recordings were made using a free Android application called Easy Voice Recorder. The original sample rate of the recordings was 44.1 kHz and noise cancellation was turned-off.

Tasks: Each participant was asked to read three texts for 5 minutes each for a total of 15 minutes. The texts were chosen from different domains to capture as much phonetic difference as possible. The source of the first text is the opening of a short child story called "De drie beren". It involved conversation between characters which could influence pitch. The second text was different sections blended from the Wikipedia page of The Netherlands and the city of Leiden. Finally, the third text was taken from a news article by NRC on climate change. The full corpus can be found in the Appendix (9.1-9.3). The order in which the participants read the text was rotated from session to session. Some participant finished reading a text before the 5-minute mark, in which cases they were asked to repeat the text from the beginning to ensure equal amount of data for each speaker. For the media-speakers, the audio was played through a JBL loudspeaker in the same room and was re-recorded with the same phones under the same conditions. Each media-speaker had 5 minutes of voice samples. The content of speech varied from speaker-to-speaker, so this could be considered free speech.

Design: The external conditions for each recording were identical: they were recorded in the same room with as little background noise as possible. One of the phones was placed on a desk close-up to the speaker to collect clean, less noisy samples. The distance corresponded to about 0.5 meters. The second phone was placed at the other end of the room, about 3 meters from the speaker on another desk. Finally, the third phone was placed inside the pocket of the speaker with the microphone facing the bottom of the pocket. The placement of the three smartphones had the purpose of capturing different conditions that can occur during real-life usage. This setup is visualized in Figure 2.

Processing: The recordings were converted into 16-bit .wav files, and all audio recordings have been down-sampled into 8, 16 and 32 kHz streams. 8 kHz is the lowest adequate frequency to represent human speech (with some disruptions to sibilants), 16 kHz is frequently used for modern Voice-Over-IP telephone systems and 32 kHz is standard for high-quality digital microphones. As mentioned in Section 3.1.2, the development set is sampled at 16 kHz, so the input data for the verification also needs to be sampled with the same rate. Thus, the resolution of the already down sampled 32 kHz dataset is lowered once more to 16 kHz. An opposite action is taken with the 8 kHz recordings which is up sampled to 16 kHz before using them for the verification. Besides these re-sampling processes, no volume normalisation, or any other adjustments to the original recordings were done. Due to the nature of the reading-task, the recordings contain almost exclusively speech segments, with some pauses between sentences. The ratio of speech to silence is ≈ 85% for the whole dataset. We also calculated the average signal-to-noise ratio for the three different conditions. The values in decibel are shown in Table 1.

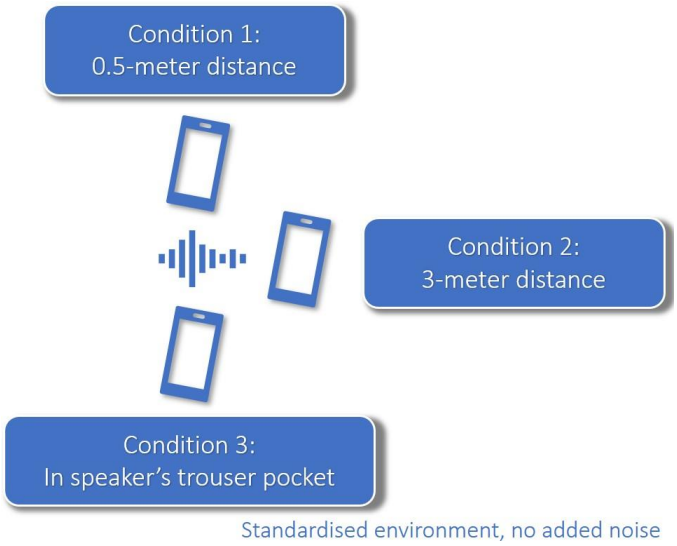


Fig. 2. Outline of the data collection setup.

Condition 1 (close-up)	Condition 2 (far away)	Condition 3 (pocket)
58.63 dB	56.49 dB	61.86 dB

Table 1. The mean Signal-to-Nosie ratio of the three conditions across all speakers shown in Decibel.

3.4 Feature extraction

As addressed in Section 2, the type of input depends on the chosen model architecture, so the feature extraction slightly differs from model to model. The models described in this work utilize two types of features: raw waveforms and 40-dimensional log mel-filterbanks (which we call MFBs). MFBs are almost equivalent to MFCCs except for the lack of the final discrete cosine transform operation to decorrelate the features. Thus, the MFCC coefficients correlate much less with each other which reduce the effectiveness of convolutional operations on them [9]. For SV based on

GMM-UBM/ i-vectors, delta and double-delta features were concatenated to MFCCs to create better acoustic features. For neural networks these features are not widely used anymore due to their ability to recognize time dependency between frames inherently [3].

Raw waveforms: From the 16bit PCM recordings, the audio is loaded using Python's *librosa* library. Voice Activity Detection (a separate algorithm to detect speech segments of an audio) is not applied to the recordings because the ratio of speech to non-speech was high (as mentioned in Section 3.1.2) The down sampling and up sampling processes explained in the previous section were also carried out using the librosa package. A waveform plot from a 2-second-long example recording is shown in the top part of Figure 7.

Log mel-filterbanks: We used the *python speech features* package to extract the MFB features [57]. The process consists of the following: First, the audio is segmented into shorter frames with overlap during which time the frequencies are said to be statistically stationary (meaning they do not change much during a short intervals). Next, the power spectrum of each frame is calculated with Short-Time Fourier Transform. To condense information, we sum the energies in different regions using filterbanks. The banks are set by the Mel scale, which is linear for low frequencies but logarithmic for high frequencies to better resemble the human hearing. Finally, we take the logarithm of the total energies for each band to get log mel-filterbanks [58]. A visualization of mel-filterbank features extracted from the same a 2-second-long example recording is shown in the bottom part of Figure 7.

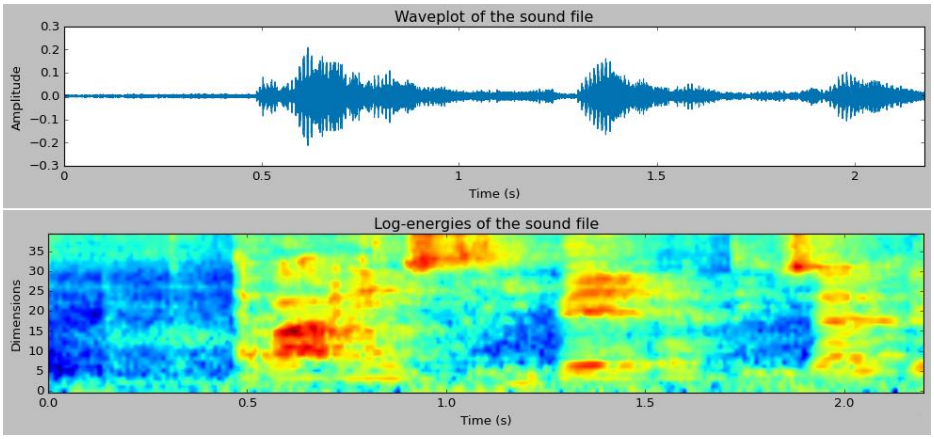


Fig. 7. Visualising the features extracted from a 3-second-long audio recording. Top: raw amplitude values. Bottom: normalised log-energy values.

3.5 Speaker Verification models

In this work we explore three different models chosen from recent literature called: RawNet2, SincNet X-Vector and Meta-Learning model. Besides their favourable performance on VoxCeleb1 trials [12], we selected a wide range of model architectures and feature input types. The implementation of all these models is openly accessible and they have been trained on the same VoxCeleb 2 dataset, so a valid comparison is possible. There were slight differences in the configurations, which will be discussed in the respective sections for the three models. A summary

detailing the most important properties of the three models can be found in Table 2.

	RawNet2	SincNet X-vector	Meta Learning
Input type	Raw waveform	Raw waveform	Log-energies
Input dim.	59049 (3.6s)	48000 (3s)	25ms x 40
Base architecture	CNN+GRU	CNN+TDNN (X-vector)	CNN (ResNet34)
Embedding dim.	1024	512	256
Loss function type	categorical	categorical	categorical + metric

Table 2. Summary of the most important characteristics of the three models compared in this paper.

3.5.1 RawNet2

RawNet is an end-to-end deep neural network with a front-end utterance-level speaker embedding extractor and a back-end classification module. The model was proposed to improve direct modelling of raw waveforms [35]. The RawNet architecture consist of CNN layers extracting frame-level representations followed by a Gated Recurrent Unit (GRU) layer which aggregates them into utterance level representations. The first layer is a Sinc-convolution which are band-pass filters constructed to only learn low and high cut-off frequencies from the data [26]. This is not only a more efficient way to discover filters but the learned filters themselves are shown to be more meaningful. The Sinc layer is followed by two residual blocks containing regular convolutional and max pooling layers with sizes 128 and 256 respectively. After each convolutional layer, Leaky ReLU activation functions are used. The following GRU initiates a self-attention mechanism which assigns weights to each frame before the aggregation process. Finally, a fully connected layer (FC) creates the 1024-dimensional speaker embeddings before the final output layer. The simplified architecture of RawNet can be seen on Figure 4. The base model achieved competitive performance with a simplified process pipeline and RawNet2 improved on it by scaling feature maps [54]. A scale vector with equal dimensions to the number of filters is used to scale the feature map both additively and multiplicatively. The resulting feature maps focus on more important features in the frame-level feature map. Compared to its predecessor, RawNet2 simplified the loss function to categorical cross-entropy which is responsible for training the contrastive representations. At runtime (both training and testing) the varying-sized input is processed-in mini-batches of 59049 samples corresponding to $\approx$ 3.6 second long 16 kHz audio. Input utterances shorter than 59049 samples are cloned to fill the remaining window. For the training of RawNet2 all 6112 speakers of VoxCeleb2 have been used by merging the training and the testing set. Pre-emphasis was not applied. We utilise the weights of the pre-trained version of RawNet2 to conduct the verification [54]. During the verification phase, the test utterances are processed with a technique called time augmentation which crops the fixed-sized input samples into multiple overlapping windows to extract multiple representations for a single input. These embeddings are then averaged to obtain one final embedding.

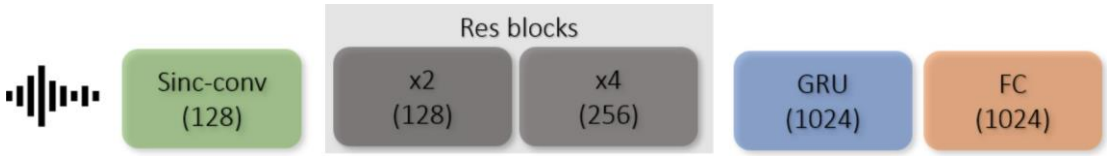


Fig. 4. Simplified architecture of the RawNet model.

3.5.2 SincNet + X-vector

The second end-to-end neural network consists of convolutional layers combined with a TDNN. More specifically, this architecture is the mixture of the standard SincNet feature extractor [26] and x-vector embeddings [23]. Again, the Sinc-convolution is placed as the crucial first layer to process the high-dimensional waveform input. Band-pass filters are implemented by parametrized Sinc functions that learn more meaningful filters compared to standard CNNs. Sinc-convolutions converge faster, have fewer parameters and are more interpretable than regular convolutions. It is followed by two regular 1D convolutional layers with kernel size 5 and max pooling layers sized 3. Then, the network connects to the X-vector architecture [23]. It consists of 5 TDNN layers followed by a statistical pooling layer before connecting to a fully connected layer. The resulting speaker embedding is 512-dimensional. In contrast to RawNet2, the model is trained using the AAM loss. For both training and evaluating the waveform input is processed with a 3s sliding window with 100ms step for which an embedding is calculated. The final speaker representation is the average of the embeddings of these frames. We utilise the *pyannote* implementation of the model with the weights pre-trained on 5994 speakers of the VoxCeleb2 training set [40].

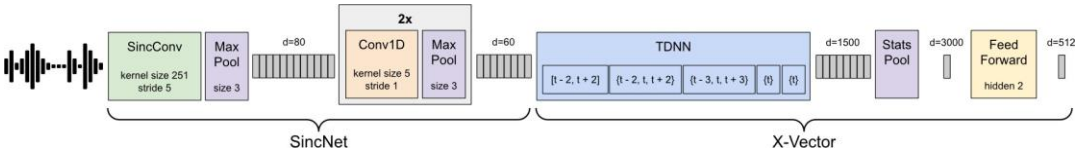


Fig. 5. Architecture of the SincNet X-vector model [40].

3.5.3 Meta-Learning

The third model utilizes a meta-learning framework and is based on the popular ResNet34 architecture [27] [30]. Meta-learning models can adapt quickly to new environments while being exposed to relatively few learning examples which is highly advantageous for SV. The added value of the residual layers with skip connections is that very deep networks can be trained. The full ResNet34 architecture is shown in Figure 6. The model uses a metric-based learning scheme which was specifically designed to perform well for short utterances of varying lengths. This is a real problem for SV where the enrolment utterances are longer than the test samples and models tend to overfit for either the short or long scenarios [55]. The length-robust model is built by simulating the enrolment/testing utterance length mismatch during training by sampling a support and a query set (a batch of training examples) from each class, where the query set have shorter variable length than the support set. Class prototypes are created by averaging the support set and enforce the query set to approximate it. Also, support and query set are classified against the whole training set. Thus, the loss function embodies both classification and metric types. The result is the reduction of class variance while embeddings are discriminate over all other classes. The model is used with

40-dimensional log mel-filterbank features calculated every 25ms with 15ms overlap. The same 5994 speakers were used for training as for the SincNet model, without applying VAD or data augmentation. Again, we utilise pre-trained weights made publicly available by the creators [30].

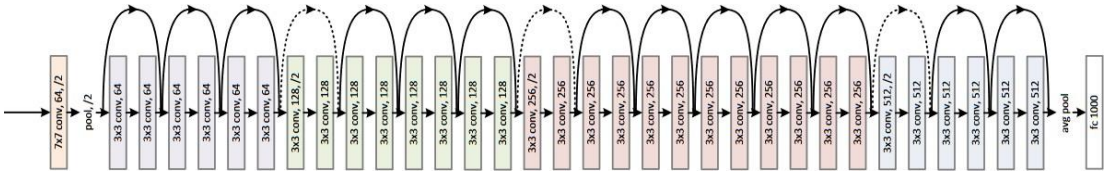


Fig. 6. Architecture of ResNet34 (on which Meta-Learning model is based) [27].

3.6 Evaluation process

The standard evaluation process for SV contains a trial which specifies which samples from the verification set are compared against one another. The baseline performance of the three models have been calculated using the VoxCeleb1 original trial [12]. For our new verification dataset, we create the trial the following way: The enrolment samples are chosen from the subject speakers' recordings of the first text and the first condition (child story, close-up scenario). The rest of the recordings (including the media-speakers) are eligible for inclusion in the test set (Table 3). To construct a trial, an equal amount of positive and negative test examples is randomly sampled for each enrolment utterance. Thus, a trial is balanced but only a small subset from the available test examples is included. To avoid reporting biased performance scores, we repeat the trial building process multiple times (each time randomly sampling from the test set) and report the average error over all the different trials. The error measure used is the Equal Error Rate, which is a standard metric for SV models.

3.6.1 Equal Error Rate

After extracting speaker embeddings in the front-end, the back-end of the model calculates the similarity score between two embeddings (enrolment vs test). The similarity score is the inverse of the cosine distance, which is the normalized dot product of the enrolment speaker embedding vector (e) and the test embedding vector (t) [59]:

$$cos(e,t) = \frac{e \cdot t}{||e|| \cdot ||t||} \tag{6}$$

The Equal Error Rate is the value which determines the threshold for which the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) is equal:

$$EER = \frac{FAR + FRR}{2}, \quad \text{if FAR} = \text{FRR} \tag{7}$$

As a rule, FRRs rise with increased sensitivity while FARs drop. The advantage of the EER metric is that it gives a measure of the overall accuracy for biometric systems without the need to manually choose the sensitivity value. In a real-life application, the enrolment speaker is already saved in the model and is compared to each candidate sample that is fed to the system outputting

a similarity score. For the scientific evaluation of models, an evaluation protocol called a trial is defined which includes candidate pairs to be compared against each other. As the test samples outnumber the enrolment samples, a trial where each sample is compared to each sample is unbalanced by default.

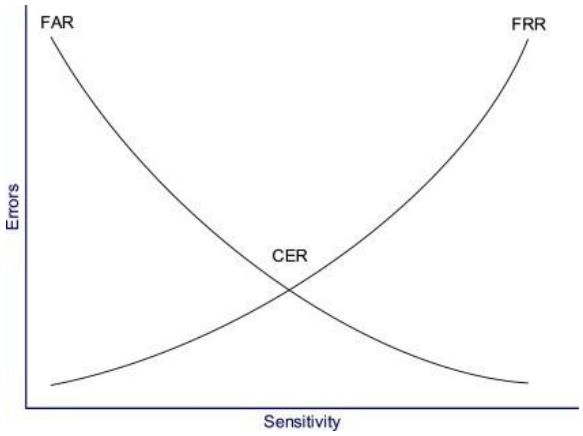


Fig. 8. The equal error rate - also known as the crossover error rate (CER) - is the intersection of the false acceptance rate (FAR) and false rejection rate (FRR) lines [60].

The number of positive and negative examples were chosen to be 100 each, and the number of trials to be 10. We selected this configuration based on multiple experiments with different values. We analysed the conversion of the standard deviation of the average error scores for each value between 1 and 200 for the number of negative/positive examples. Next, we analysed the conversion of the same measure for each value between 1 and 50 for the number of repetitions. Figure 9 shows the outcome of these experiments and the values chosen for the final configuration. As the trial size can exponentially increase with higher iterations and samples, the time of evaluation has also been taken into consideration when picking these values. Furthermore, the enrolment and test lengths were cropped to 5,10,15,30,60 and 1,2,5,10,15 seconds respectively. These lengths were chosen to provide challenging trial scenarios which appear in real life.

subject speakers	Text 1 (child story)	Text 2 (Wikipedia)	Text 3 (article)
Condition 1 (close-up)	ENROL	TEST	TEST
Condition 2 (far away)	TEST	TEST	TEST
Condition 3 (pocket)	TEST	TEST	TEST

media speakers	Free speech
Condition 1 (close-up)	TEST
Condition 2 (far away)	TEST
Condition 3 (pocket)	TEST

Table 3. Table showing the verification split of the recordings for the two speaker groups. During evaluation, only the subject speakers' samples from 'Condition 1' and Text 1 were enrolled.

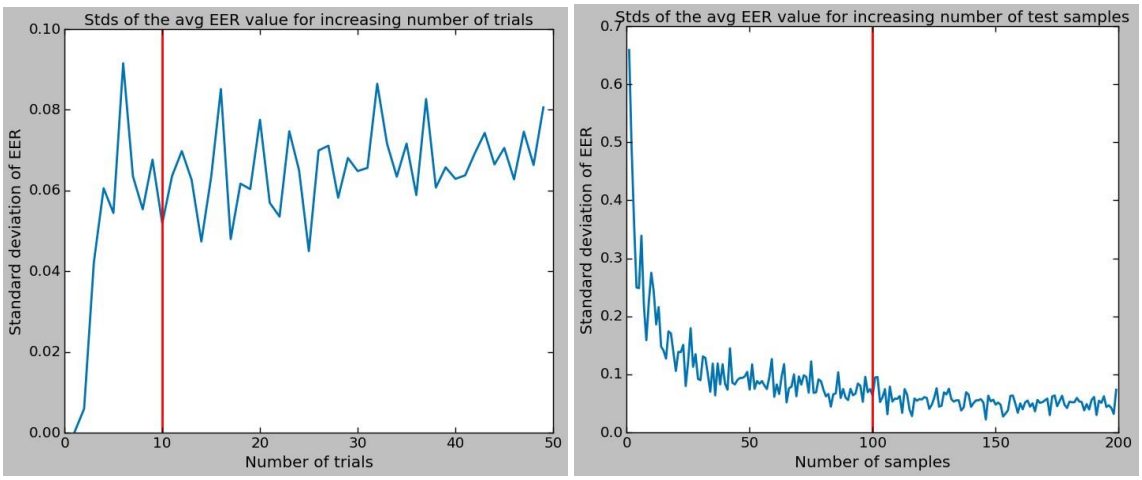


Fig. 9. Standard deviation of the average EER scores calculated for different runs using different values for the number of iterations (left) and the number of test samples(right). The red lines represent the chosen values of 10 and 100 respectively. We argue that the errors converge from these values onwards.

4. RESULTS

4.1 Overall performance

We extract embeddings for all speakers, sampling rates, and lengths (discussed in Section 3.4). For each sampling rate, the average performance is reported across all conditions. We enrol speakers according to Table 3 and apply the evaluation process discussed in Section 3.4. The average EER for the three models across the three examined sampling rates can be observed from Table 4 where the best EER scores are shown in bold for each model.

First, we observe that at 8 kHz all three models perform worse than at 16 and 32 kHz which is not surprising due to the lower original resolution of the inputs. Much improvement is not seen for two of the models (RawNet2 and Meta- Learning) when increasing the sampling rate to 32 kHz. We see that the difference between 16 kHz and 32 kHz recordings are negligible. We can relate these errors to the respective performance reported for each model in their reference papers (Table

5). Not surprisingly, the model which performed best on the VoxCeleb1 trial (Meta-Learning) performed best for our custom trial on the smartphone recordings. A discrepancy is noticeable between the performances of RawNet and SincNet: the former performed worse on our trial while SincNet showed higher errors on the VoxCeleb1 trials.

	RawNet2	SincNet X-vector	Meta-Learning
8 kHz	12.78	14.52	9.79
16 kHz	**7.59**	7.97	**4.28**
32 kHz	7.6	**7.96**	4.28

Table 4. Overall performance of the three models with different original sampling frequencies. Meta-Learning at 16 kHz and 32kHz yields the lowest EER with 4.28 followed by SincNet at 32 kHz with 7.96.

RawNet2 [54]	SincNet X-vector [40]	Meta-Learning [30]
2.48	3.5	**2.08**

Table 5. Performance of the three models reported in their original papers by the authors. The best reported EER is shown.

4.2 Performance per utterance length

We compare the models' average performances when using enrolment and test samples cropped to sizes defined in Section 3.4. From now on, only the 16 kHz performances is reported for each model. As seen in the left side of Figure 10, the EER drops when increasing both enrolment and test utterance lengths. Increasing enrolment lengths from 5 seconds to 60 seconds decreases EER on average by 3.5 for RawNet2 and SincNet. The decrease in EER for the Meta-Learning model in the same range is only about 1. RawNet2 and SincNet perform almost the same for enrolment lengths 5 and 10 seconds. For longer utterances, RawNet2 outperforms SincNet by about 0.5 EER. Meta-Learning model outperforms both for all lengths. It is also observed that the decrease in EER is not linear, the largest relative drop of EER was seen jumping from 5 seconds to 10 seconds for all three models.

On the right side of Figure 10, test length increase causes larger drops in EER. The error is almost 6 times higher at 1 second test utterances than at 15 second ones for Meta-Learning model and around 4 times higher for the other two. Again, Meta-Learning performs best overall, SincNet outperforms RawNet2 for all but one setting. Also, the steepest drop in EER is seen at the jump from 1- to 2-second-long utterances, while the decrease of the errors slows at longer utterances.

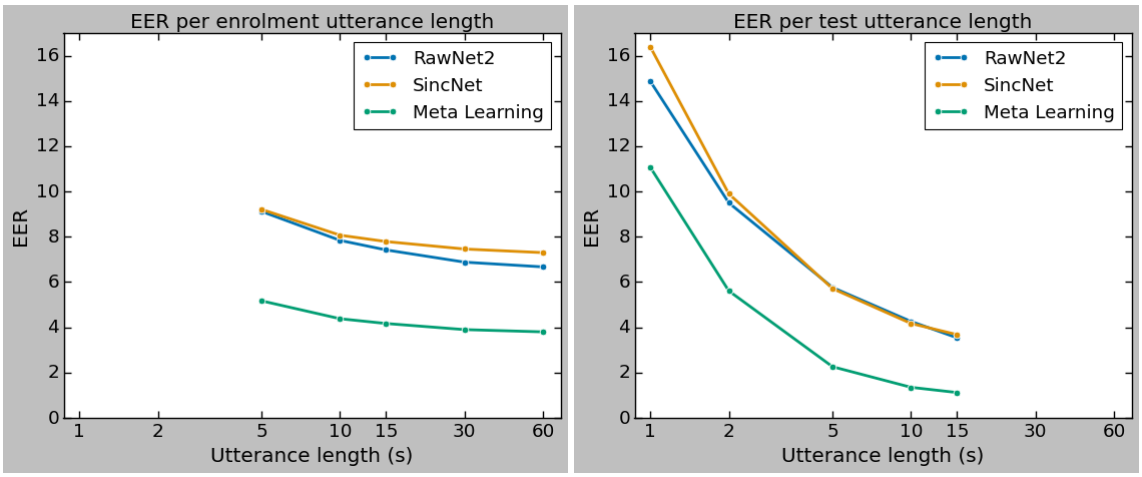


Fig. 10. EER of the models for different enrolment utterance lengths (left) and test utterance lengths (right). The performance is reported at 16 kHz original sampling rate.

4.3 Effect of recording condition on the performance

We compare the performances for the three-phone set-up to explore the effects of the different environmental conditions. We are reporting performance of the 16 kHz recordings for each model. For all the following experiments, we apply some special constraints for the reported results to reflect a more realistic setting. In the following tables the EER calculated for comparing 30-second-long enrolment utterances against the average performance for 1,2 and 5 second test utterances are reported.

Table 6 shows the results comparing the usual enrolment utterances to condition specific test utterances. Reading the EER scores from the table we see that 'Condition 1' has the lower errors. It is followed by 'Condition 3' with an increase between 1 and 2 EER. The EER for 'Condition 3' is the highest for all the models. The EER is lowest when testing only close-up recordings with Meta-Learning model (5.67) and highest when testing far-away recordings with the SincNet model (15.22).

Next, we report the results of an altered trial when enrolling pocket utterances only. The results are shown in Table 7. We only evaluated the trial 'Condition 3' vs 'Condition 3'. So, the row in Table 7 corresponds to the first row of Table 6. The errors are lower than the corresponding best scores for the previous experiment (Table 6, first row).

	RawNet2 (16 kHz)	SincNet X-vector (16 kHz)	Meta-Learning (16 kHz)
Condition 1 (close-up)	8.69	8.85	**5.67**
Condition 2 (far away)	13.62	16.47	9.42
Condition 3 (pocket)	9.83	10.51	6.79

Table 6. EER for the different recording conditions (best sampling rate scenario is reported).

	RawNet2 (16 kHz)	SincNet X-vector (16 kHz)	Meta-Learning (16 kHz)
Condition 3 (pocket)	6.27	7.65	**2.37**

Table 7. EER when enrolling 'Condition 3' utterances and testing against other 'Condition 3' samples.

4.4 Performance per gender

As mentioned in Section 1, SR systems can be affected by whether the speaker is a male or a female, so we investigated if the EER performance was influenced by the speakers' gender. In the first trial we only compare females to females and in the second trial only males to males. For this trial we only enrolled and tested the CHDR subject speakers. As we observe from Table 8 the performance drops for the female only trial. The best performing model for both the only-female and only male trial is the Meta-Learning system with 7.72 and 3.9 EER respectively. For the other two models, the difference is more than 10 EER between the male and female trial. These models also underperform Meta-Learning model in this challenging scenario.

In Table 9 we show the performance of the models on the original VoxCeleb1 trial when conducting separate trials for the two genders. Compared to our findings (Table 8), we do not observe difference between the two genders at all. We also cannot see a large performance difference between the Meta-Learning model and the other two models.

	RawNet2 (16 kHz)	SincNet X-vector (16 kHz)	Meta-Learning (16 kHz)
females vs females	22.45	22.95	7.72
males vs males	10.91	12.19	**3.9**

Table 8. EER when comparing the two genders separately.

	RawNet2 (16 kHz)	SincNet X-vector (16 kHz)	Meta-Learning (16 kHz)
females vs females	2.49	5.24	**2.88**
males vs males	3.0	5.41	3.5

Table 9. EER of the two gender trials on the VoxCeleb 1 trial.

To interpret the previous two experiments better we can visualise the embedding vectors that are extracted from the last hidden layer of the SincNet model. The 512-dimensional representations can be projected to a 2D space by running dimensionality reduction using the t-SNE algorithm. We extract 30-second-long enrolment embeddings and 5-second-long test embeddings from one of the worse performing models in this domain (SincNet) for females and males separately. Figures 11 and 12 illustrate the extracted embeddings in a 2D space for males and females respectively.

The female clusters are less separated from each other compared to males meaning that the cosine distance calculated by the models are lower causing more ambiguity. We also observe the closeness of the 'far away' condition representations (depicted with the diamond shape) meaning

that from a distance the female voice features become to merge (this phenomenon occurs for males as well with less negative effect on the overall performance). We see an emerging pattern of the 'far away' embeddings. They break away from the main cluster and approach another drifting cluster of another speaker (subjects 8, 12, 15, 16 for males and subjects 1, 9 for females).

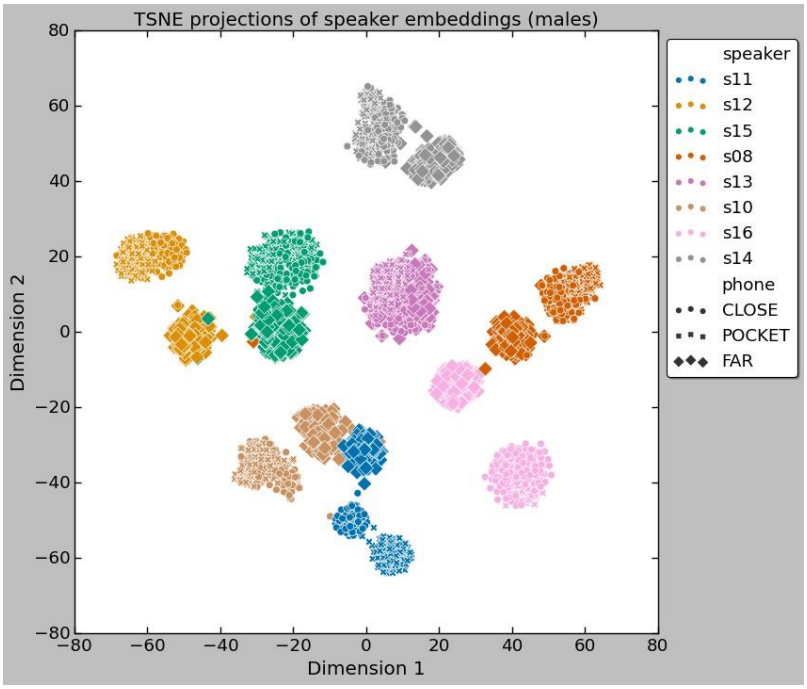


Fig. 11. 2D projections of the speaker embedding vectors of male speakers. Each colour represents 1 speaker, and the shapes represent the source phone condition.

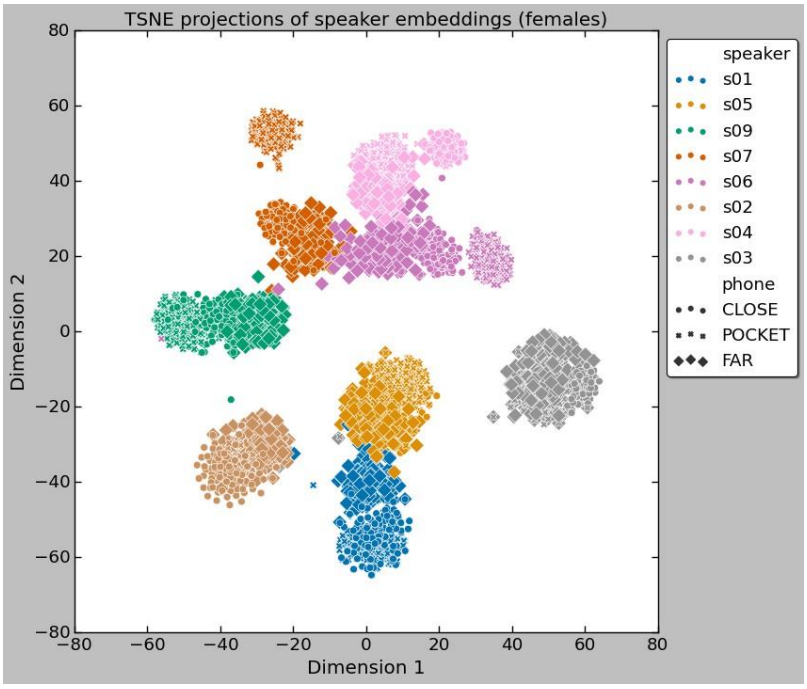


Fig. 12. 2D projections of the speaker embedding vectors of female speakers. Each colour represents 1 speaker, and the shapes represent the source phone condition.

4.5 Performance per text

Finally, we create trials containing test utterances related to a specific text. In Table 10, the three row corresponds to trials 'Text 1' vs 'Text 1', 'Text 1' vs 'Text 2' and 'Text 1' vs 'Text 3' respectively. The EER is lowest for the Meta-Learning model, ≈ 7.9 for all three texts. The EER scores for the RawNet2 is ≈ 11.7. For SincNet we observe a 1.3 EER increase from Text 3 to Text 1.

	RawNet2 (16 kHz)	SincNet X-vector (16 kHz)	Meta-Learning (16 kHz)
Text 1 (child story)	11.71	13.44	7.86
Text 2 (Wikipedia article)	11.88	12.56	7.9
Text 3 (news article)	11.6	12.17	7.94

Table 10. Performance for the different scripts read by the participants.

4.6 Computational efficiency analysis

We examined the averaged runtime of the three models on the same hardware configuration. We randomly selected one voice sample from the verification dataset and repeatedly extracted the corresponding speaker embedding 1000 times. We explored the performance for the different testing utterance lengths (1,2,5,10, and 15 seconds). Each run was timed, and the density of the timings are shown in Figures 13 and 14. For each experiment, one specific graphics processing unit (GPU) was used (GTX 970 Ti). Figure 13 shows a specific set of runs carried out with 5s long, 16 kHz utterances. The average time to create speaker representation took ≈9/10 milliseconds (ms) for RawNet2, ≈5 ms for SincNet and ≈8/9 s for Meta-Learning. The extraction time varied relatively little for the first two models with standard deviation measured lower than 1ms. Meta-Learning had a much more varying runtime with a long tail towards both quicker and slower executions. The average extraction time fell between those of the other two models. Judging from Figure 14, RawNet2 embeddings were created with the same rate independently of the input length. The speed of SincNet drops when the utterance length increases but achieves similar speed for longer utterances than Meta-Learning for short utterances. This phenomenon is observed for Meta-Learning model as well but the relative decrease in speed has a much larger magnitude (from ≈8/9 ms to ≈20 ms).

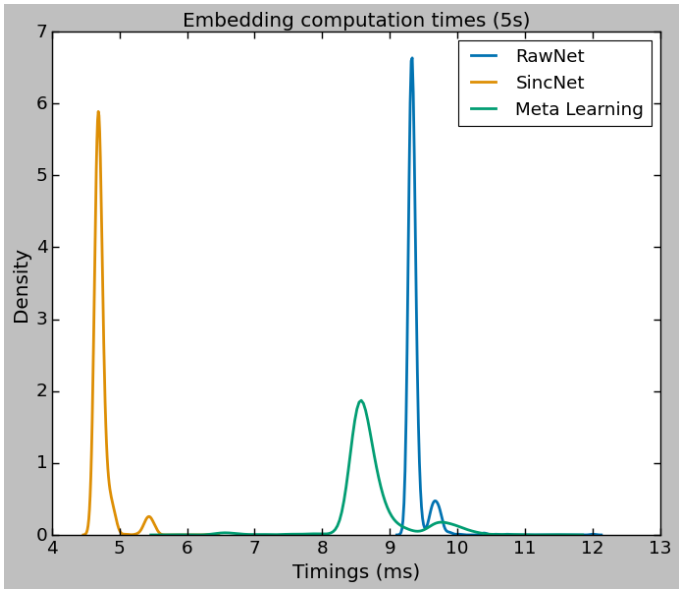
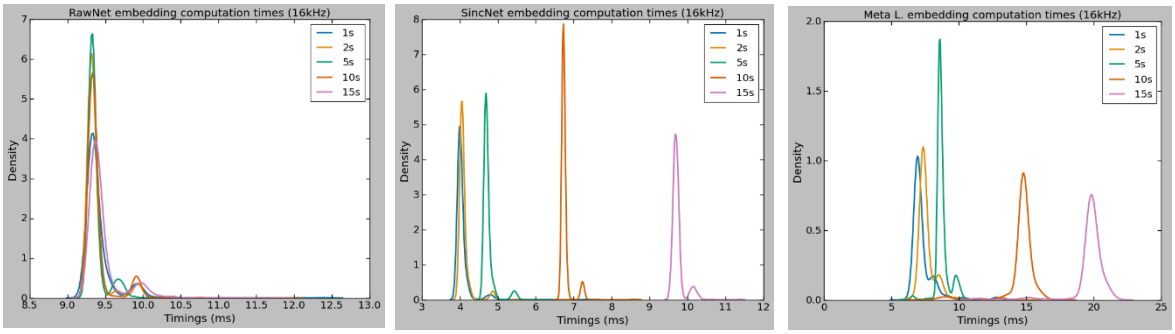


Fig. 13. Computation times of a 16kHz, 5-second-long utterance per model. Calculated by inputting and extracting the same speaker representation 1000 times.



(a) RawNet2 **(b)** SincNet **(c)** Meta-Learning

Fig. 14. Computation times of each model for different utterance lengths (1,2,5,10,15 seconds).

5. DISCUSSION

For this study, we investigated the optimal configuration and SV model for smartphone-based recordings. We consider the feasibility of this set-up successful for all three models, as they were able to extract speaker embeddings that were unique to each speaker. This statement is supported by the 2D visualisations of the embeddings in Figures 11 & 12 where speaker clusters are clearly visible even when the corresponding trial has a relatively high EER. The overall performances (Table 4) could be considered good when compared with their baseline errors (Table 5), taking into consideration that we have used three different conditions and utterance/test pair lengths as small as 5s/1s.

The performance of the models was dependent on the original sampling rate of the data. Since all three models have been trained using VoxCeleb2 sampled at 16 kHz, it was necessary to resample the original sampling rates of 8 and 32 kHz to 16 kHz when applying them in the verification. Down sampling from 32 kHz to 16 kHz had no effect whatsoever on the performance when compared with the 16 kHz trials (Table 4, first two rows). A trivial explanation is that we ended up with the same resolution when down-sampling it once from 44.1 kHz to 16 kHz and twice: first to 32 kHz and then to 16 kHz. However, a reverse operation was needed when dealing with the 8 kHz samples. Instead of down sampling, we applied up sampling on the data which added additional artificial datapoints. This had a visible effect on the performance with ERR scores dropping by 5 for RawNet and Meta-Learning and 7 for SincNet. The fixed training sampling rate affects some of the model parameters (e.g., for RawNet2 we specifically process 59049 samples as inputs) which determines the optimal input size for the models. The assessment of different sampling rates has been limited by the constraint of the fixed parameters of the pre-trained models. This could be considered an advantage since all three models performed well for 16 kHz which is the most widespread sampling rate. The models should be retrained on the same dataset down sampled to 8 kHz or up sampled to 32 kHz to assess performance. An alternative solution to manual up and down sampling would have been to use multiple phones configured to record with these sampling rates by default.

Different enrolment and testing utterance lengths have been analysed and the lowest errors were obtained for the longer utterances. This is an expected behaviour because the models simply have more data to create better representations. The problems with both the enrolment-test sample mismatch and the training-verification dataset mismatch must be taken into consideration when generalizing from these results. Previous studies showed that models tend to overfit for either long or short utterances with decreased generalizing ability for mixed-length input [61] [30]. In our experiments we have not experienced these issues, the scenario [enrolment length: 60s, test length: 15s] yielded the best results for these trials for all models. Previous research of SV models addresses the question of the effects of input sample length on model performance [59][9]. Enrolment length can also be interpreted by the volume of enrolment samples in the case of averaging enrolment representations into one embedding. In real-life SV, the enrolment process is only limited by the speaker's willingness of providing enough samples while the length of test samples encountered is boundless. For test samples the steepest drop in performance was seen from 2 seconds to 5 seconds ($\approx$ -4 EER). The small magnitude of performance drop between 30 and 60 second enrolment utterances shows that longer enrolment samples are redundant.

As noted earlier (Section 2) we expected these models to perform gender- independently. Instead, we had seen a large drop in performance for the female speakers (Table 8) for which we identified two potential causes. First, the low subject variability (only 8 female participants) in terms of age could have simply limited the reliability of the results. A more plausible reason could be the lower speech energies (volume level) of the verification dataset for the female group. This phenomenon could be explained by the different signal-to-noise ratio calculated for the female recordings (57 dB) versus the male recordings (61 dB). This is significant considering the 3dB rule: Every 3dB change represents a doubling or halving of sound energy [62]. It is possible that due to the lower overall volume level of female voices (and the higher noise level therefore) the models failed to create representative embeddings. We also justify this explanation by evaluating the performance for the two genders separately for the VoxCeleb1 official trial (Table 9). Performance on that trial is similar for females and males. Previously we were conducting more general, cross-gender trials where we included the media speakers as well. We mark that for this more challenging task, two models (RawNet2 and SincNet) struggled to yield low error rates.

One of the crucial issues with real-life remote trials is the variability of the phone location in relation to the speaker. We found that the performance of the 'close-up' condition and the 'pocket' condition is relatively small and after further investigation we found that the 'pocket' recordings provided the "cleanest" audio and the best performance (Tables 6 and 7). This is not surprising when recognizing the Signal-to-Noise ratio for each condition (Section 3.2). 'Condition 3' corresponds to a different environment that distorts the acoustic properties of speech (inside the pocket). Thus, it is surprising to see that the performance is not much worse than for 'Condition 1'. We must address the bias that is introduced into these findings by the specific verification split that was used (see Table3). All the utterances in the enrolment set were recordings corresponding to 'Condition 1' so we expect the trial ('Condition 1' vs 'Condition 1') to perform better than ('Condition 1' vs 'Condition 2') or ('Condition 1' vs 'Condition 3'). As we observed, the EER scores of 'Condition 3' had not fall behind the scores for 'Condition 1' (which was used as enrolment). Thus, we investigated this scenario further by replacing 'Condition 1' recordings in the enrolment set by 'Condition 3' recordings. We recalculated the errors for the new trials ('Condition 3' vs 'Condition 3') and the results were better than the previous trial as reported in Table 7.

We verified the 'text-independent' quality of the models by comparing performance when restricting the testing set to only include utterances for a specific text. We saw very similar performances across each text meaning that the content was not a factor for the models. Note that due to the nature of the reading- task, the performance for free-speech containing certain emotions or serious variability in pitch cannot be assessed. The same bias discussed in Section 5.4 also applies to this trial, but the performances observed for this scenario were not interesting to demand further investigation.

In terms of speed, SincNet model proved to be the fastest overall for 16 kHz, but performance drops significantly for other sampling rates. It can be argued that a few milliseconds difference between the average runtime would not be noticeable in real-life application where samples are seconds long. I recommend the use and further investigation of two models, Meta-Learning and SincNet. The former performs better while being slower while the latter compensates higher errors with speed.

5.1 Limitations

There were two main aspects where constraints had to be set to make the study feasible: the smartphone verification dataset and the SV models used. A critical aspect of speaker recognition is the intra-speaker variability, meaning that in different environmental and temporal conditions the voice characteristics might change. This session variability could not be assessed during this study due to the limited availability of the subjects. It can be presumed that some of the performances were overrated due to the short timeframe to record voice samples from a speaker (one session of 15 minutes). There were no efforts made to select subjects (both CHDR and media speakers) from all age groups, the variability for the CHDR subjects (especially the female group) was low (between 20 and 30 years old). Although the data collection protocol was created to resemble one at a future application, the standardised nature of the sessions omits some of the challenges that would occur in real-life. Due to the nature of the standardised data collection protocol, the participants were stationary during the reading activity. Thus, the pocket of their trousers created a protecting bubble that filtered noise and the models could exploit these signals to create better speaker representations. In real-life the recordings would be more prone to noise due to unpredictable movements of the body part. Due to the nature of the reading task, a limited range of emotions and pitch difference have been captured. Recent studies showed that both the emotional state of a speaker and the intentional disguise of voice (an actor playing another character) have significant effect in speech production [63,64]. The recordings contain very little background noise, and the silent parts are only taking up $\approx 15\%$ of the audio. Therefore, some pre-processing steps then were not necessary for this study might be crucial when deploying the verification model in real-life. These include volume synchronisation, noise reduction or voice activity detection. Data augmentation techniques were also applied to the training and/or verification dataset in past studies against inadequate in-domain data [65]. We also mention the lack of overlapped speakers in the recordings, which is standard for SV but still could be considered a limitation.

We selected networks that were already pre-trained so they could be considered “off the shelf” type models. The advantage of SV models is that they generalize to unknown speakers, so the computationally expensive training only needs to happen once. There are disadvantages of this property namely that the parameters are fixed, and we have limited options when processing the verification dataset. Not only is the sampling rate set during training, but the input dimension too which is tied to certain hyper-parameters. We also noted that RawNet2 model utilised more speakers during training then the other two. The model still underperformed so we could argue that adding more speakers at this magnitude has less effect on the performance. Attempts to re-train models with custom settings were made but failed due to the huge computational expenses required to train on the VoxCeleb2 dataset. We also could not estimate the language-effect of verifying Dutch speech on models trained with predominantly English speech data. Based on all the results shown in Section 4, we suspect the influence to be marginal. Finally, all three models were tested in an offline fashion and the computational efficiency analysis did not take into consideration a potential online use-case. The speed was analysed on the same hardware, but different hardware settings or network latency were not considered.

5.2 Future work

To further improve our SV models, the goal of a future study could be to assess the intra-

speaker variability better by collecting samples across multiple days or weeks from speakers. Creating a dataset of free speech samples would benefit the evaluation for different emotions and less predictable input. To further investigate the performance of SV models, a more realistic data collection protocol must be constructed with incorporating environments of actual usage. To evaluate the pocket recordings more realistically, the speakers could be asked to be physically active during recording so that the consequent noise affects would be captured. Furthermore, the models could be trained on VoxCeleb2 down sampled to 8kHz to assess the performance of a lower sampling rate better. The computational expenses of the smaller dataset could be more feasible with limited hardware. Finally, the CHDR MORE application could be incorporated in a future study to conduct online verification, by collecting recordings with the smartphone, processing the data, and running the evaluation on the smartphone in real-time.

6. CONCLUSION

In this study, we investigated three, recent deep learning-based speaker verification models (RawNet2, SincNet X-vector and ResNet34-based Meta-Learning) trained on a large set of diverse speakers (VoxCeleb2) and verified on a new dataset tailored to clinical trial application. This is part of an attempt to improve the capabilities of the CHDR MORE® application to not only recognise the presence of human speech but to be able to distinguish between the speaker of interest and others. The deep-learning models were used as feature extractors and classifiers for the subject speakers across different recording conditions (which includes the phone placed inside the pocket), gender, utterance length and sampling rate. We also created custom evaluation trials that resemble real-life scenarios better than previous trials. The overall equal error rates of the models compared with their respective baselines (RawNet: 2.48, SincNet: 3.5, Meta-learning: 2.08) were higher (RawNet: 7.58, SincNet: 7.96, Meta-learning: 4.28) but we have evaluated on more challenging scenarios. We found that the ResNet34-based Meta-Learning model with an original sampling rate of 16 kHz inputs performed best for all scenarios. It overperformed the other two models especially on the gender-trials. In terms of speed, SincNet model provided the fastest and most consistent embeddings. However, we found the Meta-learning model to be more suitable for clinical trials due to its robust performance on challenging scenarios.

The optimal SV model could be applied in many ways for remote clinical trials in the Trial@HOME® program. First, as a security layer to ensure the integrity of the voice samples by verifying if the subject using the device is the genuine subject of interest. In this use-case, both enrolment and testing conditions can be controlled and customised. Based on our findings, placing the phone inside a pocket increases the signal to noise ratio and the performance of the model. The second, more general application is the continuous recording of a subject without constraints on phone position. Applying SV in such wild, uncontrolled conditions is challenging and requires careful model-tuning and manual data cleaning. However, for this study, we were only experimenting with three standard conditions, while in real-life the phone might be placed in other different positions. As such, this study can be considered the first of many in this topic.

7. ACKNOWLEDGEMENTS

I would like a special thank you to Ahnjili ZhuParris from CHDR for her daily guidance and feedback on all aspects of the study. I acknowledge my second supervisor Robert-Jan Doll from CHDR for his input and evaluation. I would also thank all members of the Method Development team at CHDR who showed interest and provided valuable feedback during my internship at the company. Thank you to the kind staff members of CHDR who helped recruit the 16 participants to conduct the data collection. I also thank the participants for giving me their time and effort. Thank you to prof. dr. ir. David van Leeuwen for the helpful professional discussion.

8. REFERENCES

[1] Trial@home, https://chdr.nl/trialhome/ Accessed: 2021-06-25.

[2] Huang, Zili & Wang, Shuai & Yu, Kai. (2018). Angular Softmax for Short-Duration Text-independent Speaker Verification. 3623-3627. 10.21437/Interspeech.2018-1545.

[3] Bai Z, Zhang XL (2021) Speaker recognition based on deep learning: An overview. arXiv:2012.00931.

[4] Bimbot, F., Bonastre, JF., Fredouille, C. et al. A Tutorial on Text-Independent Speaker Verification. EURASIP J. Adv. Signal Process. 2004, 101962 (2004).

[5] L. Wan, Q. Wang, A. Papir and I. L. Moreno, "Generalized End-to-End Loss for Speaker Verification," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 4879-4883, doi: 10.1109/ICASSP.2018.8462665.

[6] Balian, J., Tavarone, R., Poumeyrol, M., & Coucke, A. (2021). Small footprint Text-Independent Speaker Verification for Embedded Systems. ICASSP.

[7] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland and O. Vinyals, "Speaker Diarization: A Review of Recent Research," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 2, pp. 356-370, Feb. 2012, doi: 10.1109/TASL.2011.2125954.

[8] Fujita, Yusuke & Kanda, Naoyuki & Horiguchi, Shota & Xue, Yawen & Nagamatsu, Kenji & Watanabe, Shinji. (2019). End-to-End Neural Speaker Diarization with Self-Attention. 296-303. 10.1109/ASRU46091.2019.9003959.

[9] Torfi, Amirsina & Nasrabadi, N.M. & Dawson, J. (2017). Text-Independent Speaker Verification Using 3D Convolutional Neural Networks. arXiv:1705.09422 [cs.CV]

[10] Lu, Liang & Dong, Yuan & Zhao, Xianyu & Liu, Jiqing & Wang, Haila. (2009). The effect of language factors for robust speaker recognition. Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on. 4217-4220. 10.1109/ICASSP.2009.4960559.

[11] Rohdin, Johan & Stafylakis, Themos & Silnova, Anna & Zeinali, Hossein & Burget, Lukas & Plchot, Oldrich. (2019). Speaker Verification Using End-to-end Adversarial Language Adaptation. 6006-6010. 10.1109/ICASSP.2019.8683616.

[12] Nagrani, Arsha & Chung, Joon Son & Xie, Weidi & Zisserman, Andrew. (2019). VoxCeleb: Large-scale Speaker Verification in the Wild. Computer Speech & Language. 60. 101027. 10.1016/j.csl.2019.101027.

[13] Thullier, F., Bouchard, B., & Ménélas, B. (2017). A Text-Independent Speaker Authentication System for Mobile Devices. Cryptogr., 1, 16.

[14] Brunet, Kevin & Taam, Karim & Cherrier, Estelle & Faye, Ndiaga & Rosenberger, Christophe. (2013). Speaker Recognition for Mobile User Authentication: An Android Solution. 8ème Conférence sur la Sécurité des Architectures Réseaux et Systèmes d'Information (SAR SSI)

[15] Reynolds, D., Quatieri, T., & Dunn, R.B. (2000). Speaker Verification Using Adapted Gaussian Mixture Models. Digit. Signal Process., 10, 19-41.

[16] Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. Speech Commun., 52, 12-40.

[17] W. M. Campbell, D. E. Sturim and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," in IEEE Signal Processing Letters, vol. 13, no. 5, pp. 308-311, May 2006, doi: 10.1109/LSP.2006.870086.

[18] Kenny, Patrick & Ouellet, Pierre & Dehak, Najim & Gupta, Vishwa & Dumouchel, Pierre. (2008). A Study of Interspeaker Variability in Speaker Verification. Audio, Speech, and Language Processing, IEEE Transactions on. 16. 980 - 988. 10.1109/TASL.2008.925147.

[19] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788-798, May 2011, doi: 10.1109/TASL.2010.2064307.

[20] Yuan X, Li G, Han J, Wang D, Tiankai Z (2021) Overview of the development of speaker recognition. *Journal of Physics: Confer- ence Series* 1827(1):012125.

[21] Garcia-Romero, D., & Espy-Wilson, C. (2011). Analysis of i-vector Length Normalization in Speaker Recognition Systems. INTERSPEECH.

[22] Lei, Y., Scheffer, N., Ferrer, L., & McLaren, M. (2014). A novel scheme for speaker recognition using a phonetically-aware deep neural network. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1695-1699.

[23] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5329-5333, doi: 10.1109/ICASSP.2018.8461375.

[24] Ke Chen and Ahmad Salman. 2011. Extracting speaker-specific information with a regularized Siamese deep network. In Proceedings of the 24th International Conference on Neural Information Processing Systems (NIPS'11). Curran Associates Inc., Red Hook, NY, USA, 298–306.

[25] Bhattacharya, G., Alam, J., Kenny, P. (2017) Deep Speaker Embeddings for Short-Duration Speaker Verification. Proc. Interspeech 2017, 1517-1521, DOI: 10.21437/Interspeech.2017-1575.

[26] Ravanelli, M., & Bengio, Y. (2018). Speaker Recognition from Raw Waveform with SincNet. 2018 IEEE Spoken Language Technology Workshop (SLT), 1021-1028.

[27] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[28] Chung, J.S., Nagrani, A., & Zisserman, A. (2018). VoxCeleb2: Deep Speaker Recognition. INTERSPEECH.

[29] Y. Yu, L. Fan and W. Li, "Ensemble Additive Margin Softmax for Speaker Verification," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 6046-6050, doi: 10.1109/ICASSP.2019.8683649.

[30] Kye, Seong & Jung, Youngmoon & Lee, Hae & Hwang, Sung & Kim, Hoirin. (2020). Meta-Learning for Short Utterance Speaker Recognition with Imbalance Length Pairs.

[31] Wan, Li & Wang, Quan & Papir, Alan & Moreno, Ignacio. (2018). Generalized End-to-End Loss for Speaker Verification. 4879-4883. 10.1109/ICASSP.2018.8462665.

[32] Peddinti, V., Povey, D., & Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. INTERSPEECH.

[33] Fan, Z., Li, M., Zhou, S., & Xu, B. (2020). Exploring wav2vec 2.0 on speaker verification and language identification. ArXiv, abs/2012.06185.

[34] Lin, W., & Mak, M. (2020). Wav2Spk: A Simple DNN Architecture for Learning Speaker Embeddings from Waveforms. INTERSPEECH.

[35] Bn, LeakyReLU, & Conv (2019). RawNet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification.

[36] König A, Riviere K, Linz N, Lindsay H, Elbaum J, Fabre R, Derreumaux A, Robert P: Measuring Stress in Health Professionals Over the Phone Using Automatic Speech Analysis During the COVID-19 Pandemic: Observational Pilot Study, J Med Internet Res 2021;23(4): e24191

[37] Xiang, Xu & Wang, Shuai & Huang, Houjun & Qian, Yanmin & Yu, Kai. (2019). Margin Matters: Towards More Discriminative Deep Neural Network Embeddings for Speaker Recognition. 1652-1656. 10.1109/APSIPAASC47483.2019.9023039.

[38] Deng, Jiankang & Guo, Jia & Zafeiriou, Stefanos. (2018). ArcFace: Additive Angular Margin Loss for Deep Face Recognition.

[39] Vishnyakova, Olga & van Leeuwen, David (2020): Text-Independent Speaker Recognition based on DNN

embeddings, Master's thesis, Radboud University, Nijmegen

[40] Coria, Juan & Bredin, Hervé & Ghannay, Sahar & Rosset, Sophie. (2020). A Comparison of Metric Learning Loss Functions for End-To-End Speaker Verification. 10.1007/978-3-030-59430-5_11.

[41] Jati, A. & Peri, Raghuveer & Pal, Monisankha & Park, Tae & Kumar, Naveen & Travadi, Ruchir & Georgiou, Panayiotis & Narayanan, Shrikanth. (2019). Multi-Task Discriminative Training of Hybrid DNN-TVM Model for Speaker Verification with Noisy and Far-Field Speech. 2463-2467. 10.21437/Interspeech.2019-3010.

[42] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey and S. Khudanpur, "Speaker Recognition for Multi-speaker Conversations Using X-vectors," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 5796-5800, doi: 10.1109/ICASSP.2019.8683760.

[43] Zhu, Y., Ko, T., Snyder, D., Mak, B., Povey, D. (2018) Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification. Proc. Interspeech 2018, 3573-3577, DOI: 10.21437/Interspeech.2018-1158.

[44] Y. Zhu and B. Mak, "Orthogonal Training for Text-Independent Speaker Verification," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6584-6588, doi: 10.1109/ICASSP40776.2020.9053198.

[45] Z. Wang, K. Yao, X. Li and S. Fang, "Multi-Resolution Multi-Head Attention in Deep Speaker Embedding," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6464-6468, doi: 10.1109/ICASSP40776.2020.9053217.

[46] Xie, Weidi & Nagrani, Arsha & Chung, Joon Son & Zisserman, Andrew. (2019). Utterance-level Aggregation for Speaker Recognition in the Wild. 5791-5795. 10.1109/ICASSP.2019.8683120.

[47] H. Muckenhirn, M. Magimai.-Doss and S. Marcell, "Towards Directly Modeling Raw Speech Signal for Speaker Verification Using CNNS," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 4884-4888, doi: 10.1109/ICASSP.2018.8462165.

[48] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5206-5210, doi: 10.1109/ICASSP.2015.7178964.

[49] Ravanelli, Mirco & Bengio, Y.. (2018). Learning Speaker Representations with Mutual Information.

[50] Chowdhury, L., Zunair, H., & Mohammed, N. (2020). Robust Deep Speaker Recognition: Learning Latent Representation with Joint Angular Margin Loss. Applied Sciences, 10(21), 7522. MDPI AG.

[51] Richey, C., Artigas, M., Armstrong, Z., Bartels, C., Franco, H., Graciarena, M., Lawson, A., Nandwana, M.K., Stauffer, A.R., Hout, J.V., Gamble, P., Hetherly, J., Stephenson, C., & Ni, K.S. (2018). Voices Obscured in Complex Environmental Settings (VOICES) corpus. INTERSPEECH.

[52] McLaren, M., Ferrer, L., Castan, D., Lawson, A. (2016) The 2016 Speakers in the Wild Speaker Recognition Evaluation. Proc. Interspeech 2016, 823-827.

[53] Heigold, Georg & Moreno, Ignacio & Bengio, Samy & Shazeer, Noam. (2015). End-to-End Text-Dependent Speaker Verification.

[54] Jung, J., Kim, S., Shim, H., Kim, J., & Yu, H. (2020). Improved RawNet with Feature Map Scaling for Text-Independent Speaker Verification Using Raw Waveforms. INTERSPEECH.

[55] Kanagasundaram, Ahilan & Sridharan, S & Sriram, G & Prachi, S & Fookes, C. (2019). A Study of X-vector Based Speaker Recognition on Short Utterances. 10.21437/Interspeech.2019-1891.

[56] Lawrence R. Rabiner and Ronald W. Schafer. 2007. Introduction to Digital Speech Processing. Now Publishers Inc., Hanover, MA, USA.

[57] Welcome to python speech features's documentation! https://python-speech-features.readthedocs.io/en/latest/. Accessed: 2021- 06-25.

[58] Mel frequency cepstral coefficient (mfcc) tutorial, http://practicalcryptography.com/miscellaneous/machine-learning/ guide-mel-frequency-cepstral-coefficients-mfccs/. Accessed: 2021-05-04.

[59] Arasteh, S.T. (2020). Generalized LSTM-based End-to-End Text-Independent Speaker Verification. ArXiv, abs/2011.04896.

[60] Equal error rate, https://www.sciencedirect.com/topics/computer-science/equal-error-rate. Accessed: 2021-06-27.
[61] Li, Lantian & Wang, Dong & Kang, Jiawen & Wang, Renyu & Wu, Jing & Gao, Zhendong & Chen, Xiao. (2020). A Principle Solution for Enroll-Test Mismatch in Speaker Recognition.
[62] Understanding the 3db rule, https://pulsarinstruments.com/en/post/understanding-3db-rule. Accessed: 2021-06-25.
[63] Brown, A., Huh, J., Nagrani, A., Chung, J.S., & Zisserman, A. (2021). Playing a Part: Speaker Verification at the Movies. ICASSP.
[64] Sarma, Biswajit & Das, Rohan. (2020). Emotion Invariant Speaker Embeddings for Speaker Identification with Emotional Speech.
[65] S. Shahnawazuddin, W. Ahmad, N. Adiga and A. Kumar, "In-Domain and Out-of-Domain Data Augmentation to Improve Children's Speaker Verification System in Limited Data Scenario," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7554-7558, doi: 10.1109/ICASSP40776.2020.9053891.
[66] Kajarekar, Sachin & Malayath, Narendranath. (2000). Analysis of Speaker and Channel Variability in Speech.
[67] Reynolds D. (2009) Universal Background Models. In: Li S.Z., Jain A. (eds) Encyclopedia of Biometrics. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-73003-5_197
[68] Chen, Ke. (2003). Towards better making a decision in speaker verification. Pattern Recognition. 36. 329-346. 10.1016/S0031-3203(02)00034-1.
[69] A. Kanagasundaram, D. Dean and S. Sridharan, "Improving out-domain PLDA speaker verification using unsupervised inter-dataset variability compensation approach," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 4654-4658, doi: 10.1109/ICASSP.2015.7178853.
[70] Fagherazzi G, Fischer A, Ismael M, Despotovic V: Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice. Digit Biomark 2021;5:78-88. doi: 10.1159/000515346
[71] Goudlokje en de drie beren, https://www.beleven.org/verhaal/goudlokje_en_de_drie_beren Accessed: 2021-06-25.
[72] Nederland, https://nl.wikipedia.org/wiki/Nederland Accessed: 2021-06-25.
[73] Leiden, https://nl.wikipedia.org/wiki/Leiden Accessed: 2021-06-25.
[74] Stel dat de zee opens twee meter stijgt https://www.nrc.nl/nieuws/2021/01/31/stel-dat-de-zee-opeens-twee-meter-stijgt Accessed: 2021-06-25.

9. APPENDIX

9.1 Text 1: "De drie beren"

Text 1 is a short story called "Goudlokje en de drie beren" [71].

"Er waren eens drie beren die gezellig in een eigen huis woonden, midden in een groot bos. De ene was een klein beertje, de andere was een middelmatig grote beer en de derde was een reusachtig grote beer. Zij hadden elk hun eigen papbord: een klein bord voor de kleine beer, een middelmatig groot bord voor de middelmatig grote beer en een groot bord voor de reusachtig grote beer. En elk van hen had zijn eigen stoel: een kleine stoel voor de kleine beer, een middelmatig grote stoel voor de middelmatig grote beer en een grote stoel voor de reusachtig grote beer. Zij hadden ook elk hun eigen bed: een klein bedje voor de kleine beer, een middelmatig groot bed voor de middelmatig grote beer en een groot bed voor de reusachtig grote beer. Op een dag maakten zij hun pap klaar voor het ontbijt en deden die in hun papborden. Daarna gingen ze in het bos wandelen, terwijl hun pap afkoelde, zodat zij hun mond er niet aan zouden branden. En terwijl zij wandelden, kwam een klein

meisje, dat Goudlokje heette, bij het huisje. Eerst keek zij door het raam naar binnen en daarna gluurde zij door het sleutelgat. Toen zij niemand zag, deed zij de deur open. De deur was niet op slot, omdat de beren goede beren waren die nooit iemand kwaad deden. Zij verwachtten ook niet van anderen dat zij hun kwaad zouden doen. Goudlokje deed dus de deur open en ging naar binnen. Zij was blij verrast toen zij de pap op tafel zag staan. Als zij even had nagedacht, had zij wel gewacht totdat de beren thuisgekomen waren, die haar dan misschien uitgenodigd zouden hebben mee te eten. Het waren namelijk erg aardige beren, een beetje ruw misschien, maar dat zijn alle beren, en zij waren echt hartelijk en gastvrij. Maar de pap zag er erg lekker uit en Goudlokje besloot zichzelf te bedienen. Eerst nam zij een hapje van de pap van de reusachtig grote beer en die was te heet voor haar. En toen nam zij een hapje van de pap van de middelgrote beer en die was te koud voor haar. En toen nam zij een hapje van de pap van de kleine beer en die was niet te warm en niet te koud. Goudlokje vond hem zo lekker, dat ze het hele bord leeg at. Toen ging het meisje in de stoel zitten van de reusachtig grote beer, maar die was te hard voor haar. Daarna probeerde zij de stoel van de middelgrote beer en die was te zacht voor haar. Daarom probeerde zij de stoel van de kleine beer en die was niet te hard en niet te zacht, maar net goed. Ze ging erop zitten en ze bleef zitten totdat de bodem uit de stoel zakte en zij op de grond terechtkwam. Goudlokje stond op en ging naar boven, naar de slaapkamer waar de drie beren 's nachts sliepen. Eerst ging zij op het bed liggen van de reusachtig grote beer, maar het hoofdeind was te hoog voor haar. Daarna ging zij op het bed liggen van de middelgrote beer, maar van dat bed was het voeteneind te hoog. Tenslotte ging zij naar het bedje van de kleine beer. En dat was net goed. Ze ging erin liggen, trok de dekens op en viel in een diepe slaap. De drie beren dachten op dit moment dat hun pap wel afgekoeld zou zijn. Daarom gingen zij naar huis om te ontbijten. Goudlokje had de lepel van de reusachtig grote beer rechtop in zijn pap laten staan. Toen deze dat zag, zei hij met zijn grote zware stem: "Er heeft iemand aan mijn lepel gezeten! En toen de middelgrote beer naar zijn bord keek, zag hij dat ook bij hem de lepel rechtop in de pap stond. Hij zei met zijn middelmatig zware stem: "Er heeft iemand aan mijn lepel gezeten!" Het kleine beertje keek naar zijn bord. Hij zag dat zijn pap op was. En hij zei met zijn kleine stemmetje: "Er heeft iemand aan mijn lepel gezeten en mijn pap is op!" De drie beren begrepen dat er iemand in hun huisje was geweest. Zij gingen in de kamer zoeken of die iemand er nog was. Nu had Goudlokje vergeten het harde kussen recht te leggen toen zij uit de stoel opstond van de reusachtig grote beer. "Er heeft iemand op mijn stoel gezeten!" Dat zei de reusachtig grote beer met zijn grote zware stem. Het kussen van de middelgrote beer was ingedeukt. En de middelgrote beer zei met zijn middelzware stem: "Er heeft iemand op mijn stoel gezeten!" Je weet wat Goudlokje met de derde stoel had gedaan. Dus zei het kleine beertje met een klein stemmetje: "Er heeft iemand in mijn stoel gezeten totdat de bodem eruit viel!" De drie beren vonden het toen echt nodig om verder te zoeken. Zij gingen naar boven naar hun slaapkamer. Tja, Goudlokje had de kussens van de reusachtig grote beer van hun plaats getrokken. "Er heeft iemand in mijn bed gelegen!" Zo sprak de reusachtig grote beer met zijn grote, zware stem. En de sprei van het bed van de middelgrote beer was helemaal verkreukeld. "Er heeft iemand in mijn bed gelegen!" Zo sprak de middelgrote beer met zijn middelmatig zware stem. En toen het kleine beertje bij zijn bed kwam, lag de sprei op zijn plaats en het kussen lag op zijn plaats. Maar op het kussen lag het hoofdje van Goudlokje - en dat lag helemaal niet op zijn plaats, want dat hoofd had daar helemaal niets te maken. "Er heeft iemand in mijn bed gelegen - en zij ligt er nog in!" Zo sprak het kleine beertje met zijn kleine stemmetje. Goudlokje intussen had in haar slaap de grote, zware stem gehoord van de reusachtig grote beer en de middelmatig zware stem van de middelgrote beer, maar op dezelfde manier waarop je stemmen in je droom hoort. Van het hoge, schelle stemmetje van het kleine beertje werd zij echter wakker. Meteen ging zij rechtop zitten. Toen zij de drie beren aan de

ene kant van het bed zag staan, schrok zij zó, dat zij zich aan de andere kant uit het bed liet vallen en naar het raam holde. Het raam stond open, want de drie beren waren hele nette beren, die altijd het raam openzetten als zij 's morgens waren opgestaan. Goudlokje sprong naar buiten en rende zo vlug zij kon het bos in - en zij keek niet één keer om. Wat er daarna met haar gebeurd is, weet ik niet, maar ik weet wèl dat de drie beren haar nooit hebben teruggezien."

9.2 Text 2: Wikipedia (Nederland & Leiden)

Text 2 consist of excerpts from the Dutch Wikipedia article on "The Netherlands" and "Leiden" [72, 73].

"Nederland is een van de landen binnen het Koninkrijk der Nederlanden. Nederland ligt voor het overgrote deel in het noordwesten van Europa, aan de Noordzee. Naast het Europese deel zijn er nog de drie bijzondere gemeenten in de Caribische Zee, die ook wel Caribisch Nederland worden genoemd. Europees Nederland wordt in het zuiden begrensd door België, langs de oostgrens door Duitsland en aan west- en noordzijde door de zee. De hoofdstad van Nederland is Amsterdam, de regeringszetel is Den Haag. Ruim 19% van het oppervlak bestaat uit water en een groot deel van het land en de bevolking bevindt zich onder zeeniveau. Het land wordt beschermd tegen het water door middel van een systeem van dijken en waterwerken. Door landwinning zijn polders gecreëerd. Bestuurlijk is het land verdeeld in twaalf provincies en circa 350 gemeenten. Nederland werd onafhankelijk tijdens de Tachtigjarige Oorlog, waarin de gezamenlijke Noordelijke en Zuidelijke Nederlanden tegen de Spaanse overheersing in opstand kwamen. In 1579 vormden de Noordelijke Nederlanden de Unie van Utrecht, waarmee een nieuwe politieke entiteit ontstond. Met de Acte van Verlatinghe van 1581 werd door de gewesten van die unie de onafhankelijkheid van de Republiek der Zeven Verenigde Nederlanden uitgeroepen. Deze werd vanaf 1609 bij het begin van het Twaalfjarig Bestand internationaal erkend en na de Vrede van Münster ook door Spanje. Vanaf de Franse tijd ontwikkelde Nederland zich tot een natiestaat. De Nederlandse vorst regeerde anno 1815 ook over het huidige België en Luxemburg, evenals een aantal overzeese gebieden (Nederlands-Indië, Suriname en de Nederlandse Antillen). België werd echter onafhankelijk na de Belgische Revolutie in 1830 en Luxemburg maakte zich in 1890 los van de Nederlandse Kroon. De dekolonisatie maakte in de 20e eeuw ook een einde aan de Nederlandse koloniën. Behalve de drie Caribische bijzondere gemeenten onderhouden ook de eilanden Aruba, Curaçao en Sint Maarten een hechte band met Nederland: deze vier landen vormen sinds 2010 samen het Koninkrijk der Nederlanden. Politiek is Nederland sinds de grondwetsherziening van 1848 een parlementaire democratie met een constitutionele monarchie, een staatsvorm waarbij de macht volgens de regels gedeeld wordt door de koning(in), de ministers onder wie de minister-president en de twee kamers van het parlement. Nederland was medeoprichter van onder meer de Europese Unie, de G-10, de NAVO, de Wereldhandelsorganisatie en de OESO. Met België en Luxemburg vormt het de Benelux. Den Haag speelt een belangrijke internationale rol op juridisch gebied, als locatie voor vier internationale tribunalen en Europol. In 2009 behoorde Nederland als 's werelds zevende economie naar bbp per hoofd van de bevolking tot de meest ontwikkelde landen. Het bezette in 2013 de vierde plaats in de index van de menselijke ontwikkeling. De Nederlandse economie steunt vooral op een zeer hoog ontwikkelde land- en tuinbouwsector, de dienstensector en de internationale handel, met name op de doorvoer van goederen naar Duitsland. Leiden is een stad en gemeente in het noordwesten van de Nederlandse provincie Zuid-Holland. De Oude Rijn stroomt door Leiden voordat deze – even verderop – in zee uitmondt. Leiden is het centrum van een agglomeratie en

stadsgewest met onder meer Katwijk. Leiden staat bekend als studentenstad; het heeft de oudste universiteit van Nederland. Daarnaast is het een toeristische trekpleister, dankzij landelijk bekende musea en de oude binnenstad met grachten, monumentale bouwwerken en hofjes. De bijnaam luidt de Sleutelstad, verwijzend naar de sleutels in het stadswapen. Het historische centrum wordt gevormd door de Burcht van Leiden, een motteburcht op de samenvloeiing van twee armen van de Rijn. Rond de Burcht ligt een omvangrijke grachtengordel, in totaal zijn er 88 bruggen binnen de singel. De Leidse Loper is een wandeling langs 24 historische bezienswaardigheden in de binnenstad van Leiden. Twee van de voornaamste civiele bouwwerken bevinden zich aan de Breestraat: het Stadhuis, gesierd door de breedste renaissancegevel van Nederland, en het Gemeenlandshuis van Rijnland, dat lange tijd het hoogheemraadschap van Rijnland huisvestte. Weer andere bouwwerken getuigen van de nijverheids- en handelsgeschiedenis van de stad, zoals de Waag, de Koornbrug en enkele tientallen monumentale wevershuisjes. Eén wevershuis is als museum van binnen te bezichtigen. Ook de universiteit heeft in de voorbije eeuwen een zichtbare stempel op de binnenstad gedrukt. Noemenswaardig zijn het Academiegebouw aan het Rapenburg, met daarachter de Leidse hortus botanicus en het bezoekerscentrum van de Oude Sterrewacht, en het Kamerlingh Onnes Gebouw van de rechtenfaculteit aan het Steenschuur. Leiden telt twee universiteiten en een hogeschool. De Universiteit Leiden werd opgericht in 1575 en is daarmee de oudste universiteit van Nederland. De universiteit heeft zeven faculteiten en meer dan 24.000 studenten. Het hart van de universiteit is het Academiegebouw aan het Rapenburg. De Universiteit Leiden beschikt verder onder meer over een eigen botanische tuin, een eigen observatorium en een eigen universiteitsbibliotheek. Sinds 1983 heeft ook de Amerikaanse Webster University een vestiging in Leiden. De Hogeschool Leiden is een hbo-instelling met ruim 9000 studenten gevestigd in het Leiden Bio Science Park. In dit Bio Science Park bevinden zich vestigingen van verschillende kennisinstituten, tezamen met een aantal biotechnologische bedrijven."

9.3 Text 3: nrc.nl article

Text 3 is a recent article published on nrc.nl with the title "Stel dat de zee opens twee meter stijgt" [74].

"Als de zeespiegel na 2050 sneller stijgt dan verwacht, wat moet Nederland dan doen? Er zijn drie opties, blijkt uit allerlei plannen. Het land beschermen, zoals nu, een offensieve aanpak, of de economie oostwaarts verleggen. Lang is gedacht dat de stijging van de zeespiegel aan het einde van deze eeuw echt niet meer dan een meter zou bedragen. Maar stel dat het toch meer wordt, misschien twee meter? Han Meyer: „Hoe lang kun je de Nieuwe Waterweg dan blijven uitdiepen om tankers met fossiele brandstoffen naar binnen te laten? Hoe lang hou je dat vol, als door die uitdieping het gevaar van overstromingen door extreem hoog water en het binnendringen van zout water steeds ernstiger wordt?" Han Meyer is voormalig hoogleraar stedenbouw aan de TU Delft en de belangrijkste auteur van een intrigerend voorstel: een pleidooi om de Nieuwe Waterweg, honderdvijftig jaar geleden aangelegd, ondieper te maken en het water van Rijn en Maas grotendeels een andere kant op te sturen. Dan stroomt niet langer bijna twee derde van het rivierwater via de Nieuwe Waterweg naar zee, maar via het Haringvliet. De sluizen in de dam daar zouden dan alleen bij extreem hoog water worden gesloten. Als er tegelijk sprake is van een stormvloed en veel afvoer van rivierwater, zou het water kunnen worden opgevangen in het Haringvliet en het Volkerak, de Grevelingen en de Oosterschelde. Betekent dat het einde van de Europoort? „In plaats van een monofunctionele vaarweg kan de Nieuwe Waterweg een multifunctionele riviermonding worden,

met ruimte voor zowel op de toekomst gerichte havenontwikkeling als voor meer natuurontwikkeling, meer veiligheid, minder zoutindringing en bijzondere stedelijke milieus", aldus het rapport dat mede is opgesteld door Ark Natuurontwikkeling in samenwerking met het Wereld Natuur Fonds.Meyer zegt: „De Rotterdamse haven bestaat voor een groot deel uit opslag, overslag en verwerking van fossiele brandstoffen. Moet je dat tot in lengte van dagen in stand houden? Ik zou zeggen van niet. Je kunt bovendien niet blijvend grote tankers diep het land in laten varen. Dat is net zoiets als snelwegen aanleggen die tot diep in de stad reiken. Dat moet je niet doen. Laat alle grote schepen buiten de kust aanmeren, zoals nu op de Maasvlakte gebeurt. Ook andere havensteden zoals Hamburg, Londen en Shanghai gaan die richting op." Meyers pleidooi is een van de dertien meer of minder uitgewerkte ideeën die worden bestudeerd door deskundigen van het Kennisprogramma Zeespiegelstijging om te kijken of ze nadere uitwerking verdienen. Dit programma werd anderhalf jaar geleden ingesteld door minister Cora van Nieuwenhuizen (Infrastructuur en Waterstaat, VVD) en Deltacommissaris Peter Glas, de regeringscommissaris die onder andere moet zorgen voor de bescherming tegen het water. Er kwamen steeds meer signalen dat de zeespiegelstijging misschien niet beperkt blijft tot één meter, maar vanaf 2050 kan versnellen, bijvoorbeeld door het smelten van landijs op Antarctica. „De huidige strategie van Nederland is de kust op z'n plaats te houden door zandsuppleties, en in de zeearmen houden we de zee buiten de deur met dammen en stormvloedkeringen", legt Jos van Alphen uit, nauw betrokken bij het Kennisprogramma. „Maar wat doe je als de zeespiegel maar blijft stijgen? Dan moet je steeds vaker de stormvloedkeringen sluiten. En waar laat je dan al het rivierwater, dat natuurlijk gewoon door blijft stromen? "Er zijn veel meer plannen om de zeespiegelstijging het hoofd te bieden. Kennisinstituut Deltares heeft de afgelopen jaren maar liefst honderdtachtig plannen verzameld. „Het is belangrijk dat Nederland zich verder voorbereidt op de zeespiegelstijging, ook op een mogelijke sterke versnelling. De verschillende plannen geven inspiratie", stelt Marjolijn Haasnoot, wetenschapper klimaatadaptatie en water bij de Universiteit Utrecht en kennisinstituut Deltares. Er zijn grosso modo drie manieren om met deze mogelijke stijging om te gaan, blijkt uit de dertien plannen die nu nader worden beoordeeld. De eerste manier: het huidige vasteland blijven beschermen, met aanpassingen aan het waterbeheer. Hiertoe behoort het idee van Meyer. En ook de „denkrichting" van twintig onderzoekers van Wageningen University & Research: veel meer ruimte voor de rivieren, vooral voor de IJssel, inkrimping van de landbouw, en veel meer water in en rondom groene steden. Een tweede strategie is de offensieve, zeewaarts gerichte aanpak. Hiertoe behoren plannen om land voor de Hollandse en Zeeuwse kust aan te winnen, of het maken van een nieuwe kustlijn met daartussen ruimte voor waterberging. Ook het idee van oceanograaf Sjoerd Groeskamp van het Koninklijk Nederlands Instituut voor Onderzoek der Zee, past hierbij. Hij stelt voor om zowat heel Noordwest-Europa te beschermen door een dijk te bouwen tussen Frankrijk en Engeland, en tussen Schotland en Noorwegen."

9.4 Media speakers

id	Speaker	Sex	YouTube link
1	Mark Rutte	M	https://www.youtube.com/watch?v=uE7MT8-2tT0
2	Harry Potter luisterboek	M	https://www.youtube.com/watch?v=mgVdWD6OFis
3	Marieke Lips	F	https://www.youtube.com/watch?v=qrbpr00LEa8
4	NOS op 3	F	https://www.youtube.com/watch?v=C0qCe8ZzQII
5	Arjen Lubach	M	https://www.youtube.com/watch?v=yFYxBhMiQWw
6	Gerrit Hiemstra	M	https://www.youtube.com/watch?v=33RqwsCFtyg

7	Annechien Steenhuizen	F	https://www.youtube.com/watch?v=HUwX9QkHP08
8	YOUSSEF EL JEBLI	M	https://www.youtube.com/watch?v=sXsgYjvzP6A
9	Sterre Leufkens	F	https://www.youtube.com/watch?v=nKAqeS8B0Wo
10	Koning Willem-Alexander	M	https://www.youtube.com/watch?v=SWLyu_F35W4
11	De Eetclub luisterboek	F	https://www.youtube.com/watch?v=RC7OWRzaUeQ
12	juf m NT2	F	https://www.youtube.com/watch?v=uHzqn4_vyLg
13	Michael Pilarczyk	M	https://www.youtube.com/watch?v=ztZzcx5v-XE
14	Eric Scherder	M	https://www.youtube.com/watch?v=ybZN1x3Qylk
15	Daphne de Baat	F	https://www.youtube.com/watch?v=6RpWFBgTHZU
16	Bettina Reitz-Joosse	F	https://www.youtube.com/watch?v=AYtXJsd-vQ8