From Raw Speech to Fixed Representations: A Comprehensive Evaluation of Speech Embedding Techniques

Dejan Porjazovski, Tamás Grósz, *Member, IEEE*, and Mikko Kurimo, *Senior Member, IEEE*

***Abstract*—Speech embeddings, fixed-size representations derived from raw audio data, play a crucial role in diverse machine learning applications. Despite the abundance of speech embedding techniques, selecting the most suitable one remains challenging. Existing studies often focus on intrinsic or extrinsic aspects, seldom exploring both simultaneously. Furthermore, comparing the state-of-the-art pre-trained models with prior speech embedding solutions is notably scarce in the literature. To address these gaps, we undertake a comprehensive evaluation of both small and large-scale speech embedding models, which, in our opinion, needs to incorporate both intrinsic and extrinsic assessments. The intrinsic experiments delve into the models' ability to pick speaker-related characteristics and assess their discriminative capacities, providing insights into their inherent capabilities and internal workings. Concurrently, the extrinsic experiments evaluate whether the models learned semantic cues during pre-training. The findings underscore the superior performance of the large-scale pre-trained models, albeit at an elevated computational cost. The base self-supervised models show comparable results to their large counterparts, making them a better choice for many applications. Furthermore, we show that by selecting the most crucial dimensions, the models' performance often does not suffer drastically and even improves in some cases. This research contributes valuable insights into the nuanced landscape of speech embeddings, aiding researchers and practitioners in making informed choices for various applications.**

***Index Terms*—Speech embeddings, intrinsic evaluation, extrinsic evaluation, dimension contribution.**

I. Introduction

EXTRACTING meaningful representations from raw data is an integral process in many machine learning systems. These meaningful representations, often called embeddings, can either be learned along with the main task in an end-to-end (E2E) manner [1], [2] or pre-trained separately [3], [4]. The E2E learning of embeddings is preferred since they are part of the main system, optimised along the main task. This approach might require more training data because the model needs to learn to represent the information in a meaningful way while being optimised for the target task. In low-resource scenarios, where the data is scarce, producing meaningful embeddings can be difficult [5] [6]. To alleviate that, pre-trained models can be utilised, which are usually trained on large, general datasets and then integrated into the target application. The techniques for extracting these embeddings depend on the source data, from which the meaningful information needs to be extracted, whether that is text, audio, image, or video.

Due to the abundance of different approaches for extracting speech embeddings, finding the right one for the desired task is challenging. Most of the current research on speech embeddings does not compare the proposed methods against a variety of other alternatives [7], [8], [9]. Even when they are compared against other methods, the comparison is either intrinsic or extrinsic, but seldom both [10], [11]. Additionally, there is rarely an evaluation of the proposed embeddings in another language, besides English. Depending on the way the embeddings are learned, they might perform better on certain tasks. For example, the contrastive loss, used in the Siamese networks [12], works well for the word discrimination task, so most of the models trained with that paradigm are evaluated on the word discrimination, or a similar type of task.

Our main contributions in this work are as follows.

(a) We conduct intrinsic and extrinsic experiments, assessing the ability of the speech embedding methods to encode discriminative, prosodic, and semantic features in English, Finnish, and French. (b) We conduct compressive representation experiments, restricting the models to 10% most crucial dimensions for each task. Moreover, we perform layer analysis to determine the most optimal layer for each model and task.

Regarding (a), through word and emotion discrimination tasks, we show that some models are more robust to unseen vocabularies. Moreover, despite having fewer parameters, the base self-supervised models show competitive or even better discriminative capabilities than their large counterparts.

The findings on Gender ID are in line with emotion discrimination, showcasing the effectiveness of the base self-supervised models. Furthermore, the baseline MFCC features approach the performance of the self-supervised models while having significantly fewer parameters. On the Emotion ID task, we show

Manuscript received 17 November 2023; revised 21 May 2024 and 28 June 2024; accepted 1 July 2024. Date of publication 12 July 2024; date of current version 26 July 2024. This work was supported in part by the Foundation for Aalto University Science and Technology, in part by the Academy of Finland under Project 337073 "FIN-CLARIN as a Collaborative Platform for Speech Processing", and in part by Business Finland project LAREINA under Grant 7817/31/2022. The computational resources were provided by Aalto ScienceIT. The associate editor coordinating the review of this article and approving it for publication was Dr. Naoyuki Kanda. *(Corresponding author: Dejan Porjazovski.)*

The authors are with the Department of Information and Communications Engineering, Aalto University, 02150 Espoo, Finland (e-mail: dejan.porjazovski@aalto.fi).

Digital Object Identifier 10.1109/TASLP.2024.3426301

that most models learn language-agnostic prosodic features during pre-training, transferable across languages.

Throughout the semantic assessment experiments, we show the shortcomings of the small-scale models unable to discern the semantics. Moreover, the results demonstrate that the WavLM model pre-trained on English data yields the best results for Finnish, highlighting its adaptability to new languages.

Regarding (b), we find that for simple tasks like Gender ID, restricting the models to the most crucial dimensions leads to performance improvements, indicating that many of the dimensions in the large-scale self-supervised models do not contribute to the task but add noise instead. For more complex tasks, the large and ASR fine-tuned variants show less performance drop, possibly due to better dimension separation.

By measuring the dimension overlap across languages, we provide insights into whether the same regions play a significant role in all the languages for the same task. Our findings reveal that the crucial dimensions often have low overlap across languages, suggesting that those regions depend on the data used during pre-training. By calculating the overlap between tasks, we show that the models often learn spurious correlations, resulting in a higher overlap between Gender and Intent ID tasks than between Gender and Emotion ID. The Gender and Emotion ID tasks rely on the prosody, making them more related, unlike Intent where the semantic information plays a bigger role.

II. An Overview of Audio Embedding Extraction Techniques

To date, many different techniques for extracting text embeddings have been developed. One of the most popular ones is Word2vec [3]. These embeddings exploit the distributional hypothesis of words [13], placing words with a similar meaning closer in the embedding space. With the introduction of the transformer architecture [14], more powerful models for extracting word embeddings became available. One such system is BERT, which achieved state-of-the-art results on many natural language processing (NLP) tasks [15]. The text-based embedding solutions work well for NLP but are not directly suited for audio data.

The quality of the audio signals can vary significantly based on various environmental factors (a noisy background or different microphones). Moreover, the speech can contain hesitations, repetitions, and other disfluencies, making the same utterances sound significantly different. The variation in speech makes it extremely difficult to extract same embedding vectors for identical words. Furthermore, audio signals consist of hundreds or thousands of frames, making it difficult to compress the information in a single embedding vector without too much information loss. Additionally, unlike text, audio signals do not have clear boundaries between the words. To apply similar techniques to the ones used in the text domain, the audio signal needs to be segmented into words using some segmentation method [16]. These challenges (that arise when dealing with speech data) must be addressed if we want to have meaningful speech representations which are essential in a variety of applications, such as spoken content retrieval [17], emotion recognition [18], word discrimination [19], and music recommendation systems [20].

The majority of techniques for learning speech embeddings assume that the audio signal is segmented into words. One such technique is Speech2vec [9], which employs the same self-supervised training methodologies as the ones in the Word2vec model but on audio instead of text. Another popular self-supervised technique that uses segmented words to produce speech representations is Audio2vec [8]. This technique learns speech embeddings using a sequence-to-sequence autoencoder model. These self-supervised models do not require additional labelled information but rely only on the speech signal.

Contrary to the self-supervised techniques, some models exploit the transcripts in addition to the audio information, allowing for the utilisation of different learning paradigms. One such approach is explored in [21], which employs a Siamese-based [12] convolutional neural network (CNN) that separates the same and different word pairs by a margin. A similar approach utilising a Siamese network was introduced in [22] using a stacked long short-term memory (LSTM) network [23]. This Siamese LSTM model, together with a non-uniform negative sampling technique outperformed the Siamese CNN of [21]. Further improvement was observed by additionally utilising the character information, using a multi-view RNN [24]. In [25], the authors explored a different approach for producing acoustic word embeddings. In the study, they used a CNN together with a regression layer to extract fixed-size representations. The word-based supervised techniques work well for tasks that require small local context, such as word discrimination [19] and the more general query-by-example [26]; however, they do not take into account larger changes in the audio, such as pitch variation, associated with some emotions [27].

To model longer audio segments, the authors in [18] used a temporal convolutional network [28] to extract linguistically-enhanced sentence-level audio embeddings. The sentence-level embeddings were learned using a multi-task approach of reconstructing the acoustic features and transcribing the audio. These representations proved beneficial for the automatic speech recognition (ASR) and emotion recognition tasks, outperforming the other word-level approaches.

The audio embedding techniques presented so far use relatively simple architectures, but they may require large amounts of audio data to produce satisfying results, which in low-resource scenarios is not feasible. The large pre-trained language models dominate the NLP field due to their ability to extract powerful features even when there is no data available for fine-tuning [29]. Inspired by these language models, many pre-trained variants have been developed for audio domains, utilising large-scale self-supervised training. One such model is VGGish [30], inspired by the VGG architecture [31] used in the image recognition tasks. The VGGish model was pre-trained on a large number of YouTube videos and has been successfully applied to various audio classification tasks [32], [33]. Another popular pre-trained option is Wav2vec2 [34], consisting of CNN and transformer blocks. The Wav2vec2 embeddings were successfully used in [35] for the emotion recognition task. HuBERT [36] is another popular feature-extraction model closely related to

Wav2vec2. While both models use a combination of CNN and transformer layers, the difference is in the pre-training process.

The Wav2vec2 and HuBERT pre-training objectives are primarily designed for the ASR task. To improve the performance of the self-supervised models on other speech-related tasks, the WavLM model was introduced [37]. In addition to the masked speech prediction, this model uses the speech denoising task.

These large pre-trained models achieve impressive results on various low-resource tasks, but they contain millions or even billions of parameters, making them more costly to apply in real-world applications.

III. Related Works

To date, several studies compare speech embeddings. In [10], the authors compared three different training objectives for learning speech embeddings using convolutional and recurrent models. The models were evaluated on the German and Czech languages, using acoustic word discrimination and word phonology similarity tasks. Although their study focuses on less explored languages, it only evaluates the approaches intrinsically. On the contrary, in our study, we employ intrinsic and extrinsic evaluation. Moreover, we compare more training objectives and include large pre-trained models.

In another study [38], the authors evaluated a variety of training objectives (hand-crafted, unsupervised, self-supervised, and supervised) on four languages: English, French, Xitsonga, and Mandarin. Even though this study compares a variety of training objectives, it still does not include the state-of-the-art transformer models. In their experiments, they used both intrinsic and extrinsic criteria.

A closely related study to ours is conducted in [39], where the authors examined the effectiveness of various transformer-based self-supervised models in capturing linguistic information. They assessed these models using an audio version of the GLUE benchmark [40], generated by a single-speaker text-to-speech model. However, it is important to note that their study is primarily focused on linguistic information. In contrast, our study extends the evaluation to encompass the models' capacity to capture speaker-related information and their performance when operating with reduced capacity. Furthermore, we evaluate the models on Finnish and French datasets, assessing their robustness in a cross-lingual context. Another noteworthy distinction lies in their experiments, where all utterances are generated by a single speaker, overlooking speaker variability and the influence of diverse environmental factors. While this approach suits the evaluation of linguistic capabilities, it may not fully reflect real-world use cases.

The SUPERB benchmark [41] is a popular platform focused on evaluating self-supervised representations on intrinsic and extrinsic speech-related tasks. Even though this benchmark provides a good platform for testing the pre-trained embedding representations, it focuses on the English language. In our study, besides English, we use Finnish and French speech to evaluate the embeddings on less-explored languages. Moreover, we perform a compressive representation experiment by removing the unimportant dimensions and reducing the dimensionality of the models.

To overcome the limitation of the SUPERB benchmark, which predominantly focuses on the English language, a multilingual iteration named ML-SUPERB was introduced [42]. ML-SUPERB extends its evaluation scope to encompass self-supervised models across ASR and language ID tasks spanning 143 languages. While the ML-SUPERB focuses on the ASR task, in our study, we evaluate the models on multiple intrinsic and extrinsic tasks. Moreover, we delve into experimentation aimed at reducing the dimensionality of the embeddings. Therefore, ML-SUPERB and this research complement each other, helping the users make an informed decision when selecting a model.

In [43], the authors conducted an extensive review of self-supervised models, describing their underlying methods, datasets used during pre-training, experimental settings, and results. Even though they provide results on various tasks, as stated by the authors, those results were taken from the original papers, following different fine-tuning recipes and hardware constraints. In our study, we evaluate the models under the same conditions. Similar to ML-SUPERB, we believe that this study complements our research. While [43] focuses on more theoretical aspects, describing in detail the model architectures, their training objectives and datasets used, our study focuses on evaluating those models on different intrinsic and extrinsic tasks. Together, these studies provide valuable resources for finding the appropriate model for the task without wasting computational resources on trying many variants.

IV. Data

This section provides details about the datasets used to evaluate the models intrinsically and extrinsically in English, Finnish, and French languages.

LibriSpeech [44] is a popular English audio corpus, often used in ASR, containing read audiobooks. In the experiments, we used the 360-hour clean version to train the speech embedding approaches (small-scale ones) and to evaluate them on the Gender ID task. As development and test portions, we used the official clean splits.

IEMOCAP [45] is an emotion recognition corpus, originally containing 12 hours of speech, annotated with nine emotions. To be consistent with the other research conducted on this dataset, we used only the balanced emotion classes: neutral, sadness, happiness, and anger. The dataset is recorded in five sessions. In the experiments we used the first four sessions for training and the last for testing.

SLURP [46] is a challenging English audio corpus developed for spoken language understanding. It consists of 48 hours of audio samples, where people interact with a personal assistant. The dataset contains annotations with three levels of semantics: scenario, action, and entities. In our experiments, we selected the Intent ID task, which is a combination of scenario_action, containing 93 classes. For training, development, and testing, we used the official splits provided with the dataset.

Lahjoita Puhetta (LP) [47] is a large conversational Finnish dataset. It contains over 20,000 unique speakers speaking colloquial Finnish. The recordings cover all age groups (including small children) and diverse environments. These features make the dataset suitable for evaluating the embeddings in a real-world scenario. In the experiments, we used a subset of 360 hours to train the small-scale speech embedding models. To evaluate the models on the Gender ID and Topic ID tasks, we used the official development and test splits, which do not contain speaker overlap. The topics used in the experiments are: Animal friends, Sports moments, My surroundings, Summer, The cursed COVID, Media skills, Rated R, and Nature. In the original dataset, the Media skills topic is split into three subtopics, which we merged into one. Due to the Topic ID task being challenging for the models, we increased the training data to 490 hours but kept the same development and test splits.

FESC [48] is a Finnish emotion recognition corpus containing passages narrated by five male and four female speakers. The dataset provides annotations for neutral, sadness, joy, affection, and anger emotions, prepared the same way as [49]. Since there are no official splits, we used one speaker for testing and the rest for training.

Common Voice (CV) [50] is a multilingual collection of speech primarily designed for ASR. We used in our experiments a subset of the French version 17 corpus to pre-train the small-scale models and for the Gender ID task. The subset consists of around 267 hours of French speech, from which two hours were reserved for testing, two hours for validation, and the rest for training.

CaFE [51] is a small Canadian French emotion recognition corpus consisting of six emotionally neutral sentences acted by 12 actors. The emotions presented in the dataset are sadness, happiness, anger, fear, disgust, surprise, and neutral. The emotions are portrayed with two intensities, resulting in 936 audio samples, totalling around one hour.

HealthCall30 corpus [52] contains audio interactions between customers and call centre agents in French. The goal of this task is to predict the Request ID that the customer has, whether that is related to an affiliation (label “affil”) or some other issue, such as reimbursement (label “presta”). In this study, we utilised the subset of the corpus that was introduced for the ComParE 2023 challenge [53]. Since the true labels for the test set were not available, we used half of the development set for testing. The subset of the corpus consists of 83 hours, out of which 57 are for training.

V. Speech Embedding Approaches

This section provides an overview of the small and large-scale models used in this study to extract speech embeddings.

A. Small-Scale Models

MFCC features have been a popular choice for various speech-related tasks before the emergence of the self-supervised models. Hence, as an initial embedding method, we will delve into MFCC features, establishing them as our baseline. We extracted these features using a 25 ms window and a hop length of 10, with a selection of 23 coefficients. To capture the temporal variation in speech, we added the first and second derivatives of the features (deltas). Furthermore, we appended five frames from left and right context to the features.

Audio word2vec is an audio autoencoder feature extractor, introduced in [8]. It is an encoder-decoder architecture where the encoder takes the variable-length audio sequence and compresses it to a fixed-size vector representation. The decoder then reconstructs the original audio sequence from the fixed-vector representation produced by the encoder. The advantages of this approach are that it is fast to train and no labelled data is required. Following the original implementation, the encoder and the decoder consist of a one-layer LSTM, which in the encoder is bidirectional. Additionally, the decoder is augmented with a peephole connection [54]. During training, we applied zero masking [55], which randomly turns off elements of the input sequence with a 30% probability.

Speech2vec is a sequence-to-sequence model introduced in [9]. The training scheme is similar to the one used in Word2vec. The authors experimented with the skip-gram and CBOW techniques and found that the skip-gram technique constantly outperforms the CBOW. Due to that, in our experiments, we implemented the Speech2vec model using skip-gram training. The model consists of a one-layer bidirectional LSTM encoder and a unidirectional LSTM decoder. Following the original implementation, we conditioned each decoding step to the last hidden state of the encoder, as in [56].

Siamese neural network produces speech embeddings for three inputs, from which two have the same label, and one has a different one, referred to as a negative sample. Then, the cosine distance between the embeddings is calculated. The training is done using a contrastive loss, where the model learns to separate the same and different word pairs by a margin. In our case, the Siamese network consists of a one-layer bidirectional LSTM network. For choosing the negative sample, we opted for a simple solution of picking a random word different from the other two.

Linguistically enhanced embeddings (LEE) approach, introduced in [57], follows a multi-task training scheme consisting of a shared audio encoder and two decoders. The encoder is a one-layer bidirectional LSTM network taking speech features as input and producing fixed-size speech embeddings. The speech embeddings are then passed to the two decoders. The first decoder is an attention-based LSTM network that reconstructs the original speech features by minimising the L1 loss. Since the dimensions of the speech embeddings and the speech features do not match, we used a linear layer to downscale the embeddings. The second decoder consists of a dropout and two linear layers with a ReLU non-linearity. Its job is to minimise the L1 loss between the speech and word embeddings produced by the BERT model. In this case, the speech embeddings are upscaled using a linear layer to match the dimension of the word embeddings. Finally, the model is trained by combining both L1 loss functions with equal contribution.

B. Large-Scale Models

The large-scale models are typically pre-trained on big amounts of audio in an unsupervised way. Once pre-trained, they can be used as feature-extractors or fine-tuned for a specific task. In this study, we will evaluate the most popular pre-trained self-supervised models: Wav2vec2, HuBERT, and WavLM. Besides the self-supervised models, we additionally include the Whisper model in the evaluation to see how it differs from the rest.

Wav2vec2 is a self-supervised model, pre-trained by creating quantized representations out of the feature encoder outputs and masking parts of the feature encoder timesteps. The objective is contrastive, requiring the model to identify the right quantized representation for a masked timestep. The model consists of convolutional feature encoder and transformer layers.

HuBERT follows the same architecture as Wav2vec2, with the difference being the pre-training objective, which uses masked prediction. To generate the labels, k-means clustering is applied, which in the first iteration uses MFCC features and for the subsequent ones, the latent features extracted from the model.

WavLM is another self-supervised variant that follows a similar architecture. Besides the convolutional feature encoder and transformer layers, WavLM additionally incorporates gated relative position bias [58] which improves the ASR performance without too much parameter increase. The model is pre-trained with a masked prediction, along with a denoising objective.

Whisper is primarily made for multilingual ASR and is based on the encoder-decoder transformer architecture [14]. The pre-training of the Whisper model incorporates multi-task learning where along with the standard ASR task, the model simultaneously learns to perform translation and language identification, among other tasks. In our experiments, we used the encoder part of the model to extract the speech embeddings.

VI. Evaluation Tasks

A. Emotion and Gender ID

The goal of these intrinsic tasks is to assess the capability of the embedding methods to capture prosodic information from the speakers. Using the prosodic information is beneficial in many paralinguistic tasks [59], [60]. To this end, we evaluate the models on the Gender ID and the more challenging Emotion ID task.

B. Intent, Topic and Request ID

The objective of the Intent, Topic, and Request ID experiments is to evaluate the effectiveness of the explored speech embeddings in modelling the semantic information. To assess the classification performance of the speech embeddings, we leveraged the SLURP dataset for Intent ID, LP for Topic ID, and HealthCall for Request ID. In the English experiments, we used the 93 intent labels. In Finnish, we utilised the eight topics available in the LP corpus, whereas, in French, we used the two labels associated with the request type. These classification experiments allow us to assess the adaptability and robustness of the speech embeddings across different languages and tasks.

C. Word and Emotion Discrimination

To measure the discriminative performance of the small-scale embedding models, we employed the word discrimination task. The goal of the task is, given a set of three samples (in our case words), out of which two are from the same class (same word) and one from a different class (negative sample), to recognise which sample is from the different class. The elements in the triplet are defined as follows: the first is an anchor, the second is a positive sample from the same class as the anchor, and the third is a negative sample from a different class. The classification relies on the cosine similarity between the embeddings generated by the audio encoders for the anchor and the positive sample, as well as the anchor and the negative sample. To make the task more challenging, instead of randomly choosing the negative sample, we opted for the most similar word (by edit distance) to the other two.

For evaluating the discriminative capabilities of the large-scale transformer models, we constructed an emotion discrimination task following a similar approach to word discrimination. In this task, a set of three utterances is provided, with two sharing the same emotion label (anchor and positive) and one featuring a different emotion (negative). The objective is to identify which sample is the negative one. The comparison involves calculating the cosine similarity between the embeddings produced for the anchor and positive, as well as the anchor and negative samples.

D. Compressive Representation Assessment

Given the substantial number of dimensions in the large-scale pre-trained models (768 in the base and 1024 in the large version), we evaluated the models using only 10% of their most important embedding dimensions. We chose 10% due to the exponential distribution of the attributions, where the majority of the contributions come from those 10% embedding neurons (see Fig. 4). To identify these crucial dimensions, we employed the Integrated Gradients method [61]. This method assigns attributions to the input, in our case, speech embeddings, with respect to the true labels. We select the crucial dimensions based on the highest absolute attribution scores, indicating their significant influence on the predictions.

E. Inference Time Assessment

Given that the small and large-scale models differ in their number of parameters, we expect that to reflect in their inference time. By measuring the inference time, we provide insights into the trade-offs between model parameters/performance and computation cost. While large models often result in performance increase, they are inherently slow and often not applicable in real-world systems.

VII. Experiments

Most of the small-scale speech embedding training techniques assume that the utterances also contain word boundary timestamps. To segment the English datasets into words, we used the Wav2vec2 model, fine-tuned on 960 hours of LibriSpeech. For the Finnish experiments conducted on the LP dataset, we re-used the already validated word-level segmentation using a conventional ASR model [47] trained with the Kaldi toolkit [62]. For the Finnish FESC and the French data, we used the large Whisper model to first generate transcripts, then do the word-level alignment. As evaluation metrics for the classification tasks, we chose the commonly used micro F1 and unweighted average recall (UAR). For the datasets having official training, development, and test splits, we optimised the models on the development set and tested them on the test set. Due to clarity and simplicity, we only provide the results obtained on the test sets since in most cases, the development and test results follow the same trend.

As input to the small-scale speech embedding models, we used MFCC features with 13 coefficients, whereas the large-scale pre-trained models use raw audio as input. To make the small-scale embedding models comparable, we used one-layer bidirectional LSTM audio encoders. The implementation of the decoders differs depending on the training approach. All the small-scale speech embedding models produce 128-dimensional speech embeddings. The size of the large-scale pre-trained models is either 768 or 1024, depending on the model size.

For the classification experiments, we developed different models based on the task. For the Gender ID task (male/female), we used one linear layer that takes the speech embeddings and produces an output between 0-1. We trained the models by optimising the binary cross-entropy loss. We opted for a simple model to see whether the speech embeddings contain gender information without being specifically trained for that.

For the Emotion, Intent, Topic, and Request ID tasks, on the other hand, we used a bigger and more complex model. The network takes the speech embeddings (produced either by the small-scale audio encoders or the large-scale transformer models) as input and processes them through four bidirectional LSTM layers with 512 neurons (per direction), followed by a dropout with 30% probability. The features are further transformed by three linear layers with 1024 neurons and a ReLU activation between them. In the end, the features are processed through the output layer that produces class probabilities. We chose this architecture based on internal experiments where the goal was to make a reasonably-sized model capable of learning the tasks. We want to note here that a larger model would probably yield better results, but the idea in this study is to evaluate the embeddings under the same conditions instead of aiming to achieve the best classification results.

Due to the abundance of different large pre-trained models, we chose the most popular ones. More specifically, we chose the base, large, and ASR fine-tuned versions of Wav2vec2, HuBERT, and WavLM, as well as the small Whisper model. Some of these models are pre-trained only on English data and as of the time of writing, do not have multilingual or Finnish/French versions. Due to that, for the Finnish and French experiments, in the cases where we did not find a language-specific or multilingual model, we used the one trained on English data. These models are marked with "*" in Table II. For a detailed list of the models used and the code for reproducing the experiments, refer to our code repository.[1]

To identify the most suitable layer for each model and task, we conducted training on a subset of the original training sets, utilising every other layer in the base models and every third layer in the large versions. We opted for using a random subset, instead of the whole training set, due to the large number of models that need to be evaluated for each language and task. In the Finnish and French experiments where models pre-trained on English were employed, we selected the most optimal layer based on the outcomes of the English task. For more details about the layer analysis experiments, refer to the Figs. 9, 10, and 11 in the Appendix. We would like to note that some studies use a weighted sum [41] or average [35] of all the layers, which could potentially improve the performance. However, in this study, we chose to investigate individual layers. This approach allows us

TABLE I
GENDER ID RESULTS ON THE ENGLISH, FINNISH, AND FRENCH DATASETS

	Gen-En		**Gen-Fi**		**Gen-Fr**	
Embeddings	F1	UAR	F1	UAR	F1	UAR
Small-scale models						
MFCC	93.8	93.9	84.5	83.9	90.1	87.4
Audio word2vec	58.5	58.6	62.7	61.1	78.5	50.0
Speech2vec	**75.8**	**75.8**	67.6	66.3	**79.8**	**61.9**
Siamese	66.3	66.3	**72.0**	**71.4**	78.2	52.4
LEE	70.6	70.7	64.5	63.6	78.2	51.5
Base models						
Wav2vec2-B	97.0	97.0	**90.4**	**89.8**	95.2	92.8
HuBERT-B	97.2	97.2	89.2	88.6	95.4	92.8
WavLM-B+	96.2	96.3	90.2	89.6	96.3	95.2
Whisper-small	**98.8**	**98.8**	89.9	89.3	**96.7**	**95.4**
Large models						
Wav2vec2-L	**98.3**	**98.4**	89.5	88.9	95.0	94.1
HuBERT-L	98.2	98.2	90.0	89.4	**96.1**	**95.1**
WavLM-L	95.6	95.8	**90.7**	**90.1**	95.9	94.6
Large ASR fine-tuned models						
Wav2vec2-FT	95.7	95.8	**90.4**	**89.8**	**96.4**	95.3
HuBERT-FT	**97.6**	**97.7**	89.4	88.8	96.3	**95.5**
WavLM-FT	97.2	97.2	89.9	89.3	95.9	95.2

The bolded values indicate the highest scores achieved for a particular task and model group.

TABLE II
LAYERS USED TO EXTRACT THE EMBEDDINGS FOR ENGLISH, FINNISH, AND FRENCH, RESPECTIVELY

	Gender			**Emotion**			**Int / Top / Req**		
Embeddings	En	Fi	Fr	En	Fi	Fr	En	Fi	Fr
Base models									
Wav2vec2-B	3	3*	1	1	1*	3	7	7*	9
HuBERT-B	1	1*	1*	11	11*	11*	7	7*	7*
WavLM-B+	5	5*	5*	5	5*	5*	9	9*	9*
Whisper-sm	7	7	7	1	1	1	11	11	11
Large models									
Wav2vec2-L	4	4	4	10	16	7	16	16	4
HuBERT-L	4	4*	4*	7	7*	7*	19	19*	19*
WavLM-L	7	7*	7*	4	4*	4*	16	16*	16*
Large ASR fine-tuned models									
Wav2vec2-FT	1	10	1	1	7	13	13	16	10
HuBERT-FT	1	1*	4	4	4*	4	19	19*	10
WavLM-FT	10	10*	4	13	13*	4	19	19*	19

The "*" indicates that there was no multilingual or language-specific model. The base models have 12, while the large and ASR-fine-tuned versions have 24 transformer layers.

[1] https://github.com/aalto-speech/evaluation_of_speech_embedding_methods

to gain a deeper understanding of whether similar information is stored in the same or different layers across models.

VIII. Results and Discussion

A. Gender and Emotion ID Results

Table I depicts the results obtained on the Gender ID task. On the English LibriSpeech dataset, the small-scale models exhibit varying performances, with Speech2vec embeddings obtaining the highest F1 and UAR scores, followed by LEE, Siamese, and Audio word2vec models. In Finnish, the Siamese model achieved the best results, while in French, the Speech2vec was again the most optimal one. From these findings we can say that Speech2vec is the most robust among the small-scale solutions on the Gender ID task.

Shifting the focus to the large-scale self-supervised models, we can see small performance variations between the base, large, and fine-tuned versions. These findings suggest that regardless of their scale or additional ASR fine-tuning, all self-supervised models investigated in this research have effectively captured gender-related information, even across languages not encountered during pre-training. While the MFCC features demonstrate slightly inferior performance compared to the self-supervised models, they remain a viable option, particularly due to their light computational requirements.

It is worth noting that even though the MFCC and the large-scale self-supervised models achieve better results than the small-scale ones, they are not directly comparable. While the former ones process only the first 5 seconds of the utterance to make a prediction, the latter, small-scale models operate on a word level, therefore extracting labels for each word.

A closer examination of the layer analysis, as outlined in Table II, reveals a consistent trend that the lower transformer layers are most optimal for this task. This pattern suggests that crucial prosodic cues for gender identification are predominantly stored within these layers. These findings are in line with the Wav2vec2 layer analysis conducted in [63].

Given the apparent ease of the Gender ID task, as emphasised by the results, we extended our evaluation to assess the models' capacity to capture prosodic information through Emotion ID, presented in Table III.

From the small-scale models, the LEE approach emerges as best on English, while its performance is second-best for Finnish and French. The best results for Finnish and French were obtained by the Audio word2vec and Siamese models, respectively. Unlike the English and Finnish results, there is a considerable performance difference between the models in French when the amount of training data is small. The reason for the low performance of the Audio word2vec and Speech2vec models indicates that they require more data to learn the prosody. Compared to the baseline MFCC features, the small-scale models fall behind, but the difference is less pronounced.

Among the base pre-trained models, HuBERT stands out with the highest F1 and UAR scores for English and Finnish. However, for French, the WavLM model achieved the best score by a large margin, indicating that it is better suited for tasks with limited training data. Notably, the performance of the Whisper model falls behind, compared to the other models, indicating that it has difficulties with learning the prosody.

TABLE III
Emotion ID Results on the English, Finnish, and French Datasets

	Emo-En		Emo-Fi		Emo-Fr	
Embeddings	F1	UAR	F1	UAR	F1	UAR
Small-scale models						
MFCC	54.0	50.7	54.3	48.9	34.6	36.3
Audio word2vec	33.0	25.6	**40.2**	**27.1**	22.4	20.8
Speech2vec	30.9	25.0	39.1	25.7	20.5	19.0
Siamese	33.4	25.3	38.3	25.0	**32.7**	**33.9**
LEE	**36.3**	**25.7**	39.6	26.2	30.1	28.0
Base models						
Wav2vec2-B	60.0	57.0	45.0	42.9	37.8	38.1
HuBERT-B	**68.1**	**67.1**	**87.8**	**76.7**	31.4	29.2
WavLM-B+	66.8	62.8	84.6	74.4	**45.5**	**46.4**
Whisper-small	48.9	50.1	40.7	27.0	19.9	19.1
Large models						
Wav2vec2-L	**66.7**	64.7	82.5	63.5	51.9	**54.8**
HuBERT-L	64.5	65.0	**92.3**	**79.9**	53.2	54.2
WavLM-L	65.0	**65.5**	76.6	59.8	**53.4**	53.0
Large ASR fine-tuned models						
Wav2vec2-FT	60.9	60.8	79.3	61.8	50.6	54.2
HuBERT-FT	63.7	**65.0**	70.2	61.0	47.4	47.6
WavLM-FT	**66.9**	64.9	**87.2**	**71.9**	**55.8**	**55.4**

The bolded values indicate the highest scores achieved for a particular task and model group.

From the large versions, the Wav2vec2 model gave the best F1 score but the worst UAR on the English data. In Finnish, however, the HuBERT model outperformed the others by a large margin. In French, the WavLM gave the best F1, while Wav2vec2 gave the best UAR score. Comparing the large and ASR fine-tuned models, we can notice a pattern of performance drop in F1 when using the fine-tuned versions of Wav2vec2 and HuBERT, aligning with the observations in [64]. On the contrary, by ASR fine-tuning, the WavLM improved its F1 performance on all the datasets.

An interesting finding from the Emotion ID results is that, for Finnish, the best-performing model is the large HuBERT, which does not have a Finnish or multilingual version, indicating that the model has learned universal prosodic information that is transferable across languages. Similarly, for French, the English version of the large HuBERT and WavLM models outperformed the French-pre-trained Wav2vec2 in the F1 score.

The layer-analysis, presented in Table II, reveals that lower layers exhibit superior performance on this task, suggesting the concentration of prosodic information in these layers. However, exceptions exist, such as the base HuBERT or the English and Finnish ASR fine-tuned WavLM models.

B. Intent, Topic, and Request ID Results

Table IV shows the performance on Intent, Topic, and Request ID in English, Finnish, and French, respectively.

TABLE IV
INTENT, TOPIC, AND REQUEST ID RESULTS ON THE ENGLISH, FINNISH, AND FRENCH DATASETS

	Int-En		Top-Fi		Req-Fr	
Embeddings	F1	UAR	F1	UAR	F1	UAR
Small-scale models						
MFCC	28.7	17.6	28.6	27.2	60.0	60.1
Audio word2vec	29.1	17.9	21.2	16.4	45.8	50.2
Speech2vec	22.3	12.3	14.8	12.8	49.1	50.0
Siamese	42.3	28.1	42.0	35.6	49.0	49.8
LEE	**44.5**	**28.9**	**44.2**	**37.8**	**50.6**	**51.3**
Base models						
Wav2vec2-B	61.3	43.3	77.3	67.5	74.1	74.1
HuBERT-b	63.0	46.8	79.7	72.0	69.2	69.2
WavLM-B+	**70.4**	**50.1**	**83.0**	**74.1**	73.2	73.0
Whisper-small	55.9	36.5	82.8	72.4	**78.1**	**78.1**
Large models						
Wav2vec2-L	65.7	46.2	71.7	60.2	60.5	60.4
HuBERT-L	63.2	43.9	64.7	53.9	61.7	62.0
WavLM-L	**72.4**	**51.6**	**79.9**	**69.6**	**69.3**	**69.6**
Large ASR fine-tuned models						
Wav2vec2-FT	70.9	51.0	**80.7**	**70.6**	76.1	76.2
HuBERT-FT	**72.6**	**52.9**	79.9	69.5	70.2	70.3
WavLM-FT	68.0	48.6	73.2	62.2	**77.2**	**77.3**

The bolded values indicate the highest scores achieved for a particular task and model group.

TABLE V
ACCURACY SCORES FOR THE WORD DISCRIMINATION TASK WITH MATCHED AND MISMATCHED VOCABULARIES

	Matched vocabs			Mismatched vocabs	
Embeddings	**LibriSpeech**	**LP**	**CV**	**LibriSpeech**	**LP**
Audio word2vec	75.1	60.4	55.9	70.1	60.7
Speech2vec	66.0	62.2	53.1	66.9	59.2
Siamese	**91.0**	**74.3**	**66.0**	81.0	**69.0**
LEE	88.2	71.7	61.2	**81.9**	65.6

[English: 3715, Finnish: 4192, French: 1112 samples].
The bolded values indicate the highest scores achieved for a particular task and model group.

TABLE VI
VOCABULARY STATISTICS FOR THE TRAINING SETS

Word appearance	LibriSpeech-En	LP-Fi	CV-Fr
Unique words	59661	168103	348322
Frequency $= 1$	20315	107865	294758
Frequency > 1000	358	222	153

From the small-scale models, LEE demonstrated the best F1 and UAR scores in all the languages. For Intent and Topic ID, the Speech2vec and Audio word2vec underperformed, indicating that they are not well suited for more complex tasks. Compared to the baseline MFCC features on the Intent ID task, all the models, except for Speech2vec, achieved better scores, showcasing the need for better speech representations. We observed a similar trend on the Topic ID task, where LEE and Siamese models outperformed the MFCC features by a large margin. However, on the French Request ID task, the MFCC features are the better choice, whereas the rest are close to random guess.

From the base self-supervised models, WavLM achieved the highest F1 and UAR scores on Intent and Topic ID tasks, while Whisper was most suited for Requests ID. The impressive results that the WavLM model got on the Finnish Topic ID again demonstrate the ability of the model to generalise to unseen languages.

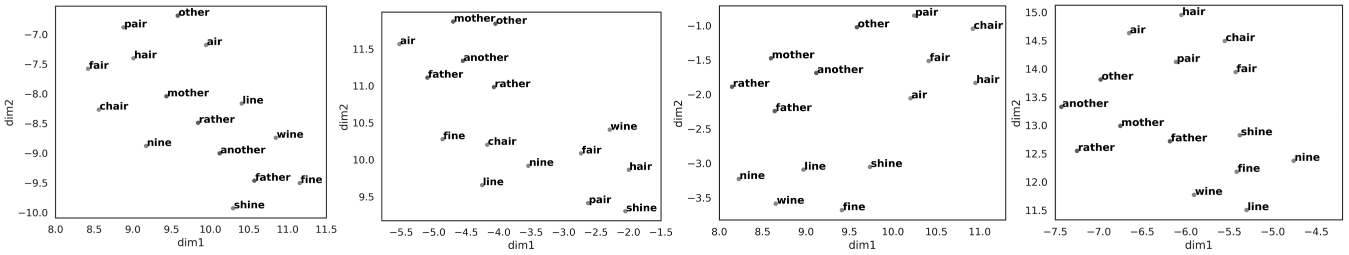
The large models exhibited a similar trend, with WavLM consistently providing the highest F1 and UAR scores. Furthermore, comparison between large and fine-tuned versions often revealed performance improvements with fine-tuning. Unlike on the Emotion ID task, where the WavLM model showed benefits from ASR fine-tuning, on the Intent and Topic ID tasks, its performance degraded noticeably. This finding shows that ASR fine-tuning the WavLM model helps with modelling the prosodic information but at a cost of losing the semantics.

Overall, these results underscore that despite the complexity of the tasks, smaller base versions often match or outperform their larger counterparts, making them a preferred choice. Layer-analysis results indicate that, for these tasks, the higher layers are generally more optimal, except for the large Wav2vec2 model on the Request ID task. These findings slightly deviate from the observations made in [65], where they found that the intermediate layers of the Wav2vec2 model encode the phonetic information. For the models pre-trained to predict discrete units, however, the phonetic information is concentrated in the higher layer, similar to our observations.

C. Word and Emotion Discrimination Results

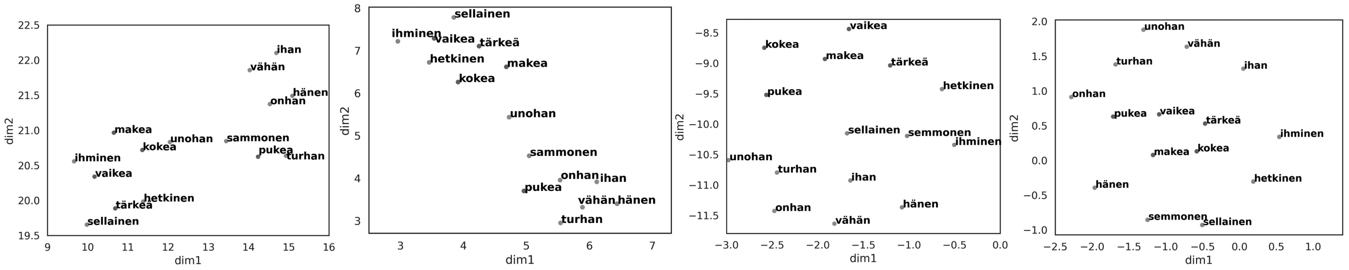
Conducting discriminative evaluation provides deeper insights into the information embedded within the models. Table V illustrates the outcomes of the word discrimination task, revealing a notable performance disparity between the English, Finnish, and French datasets. This discrepancy could be attributed to the increased difficulty of dialectical speech in diverse environments within the Finnish and French datasets. Similar to the Intent and Topic ID tasks, the Audio word2vec and Speech2vec models exhibit sub-optimal results, particularly on the CV dataset. As anticipated, the Siamese network attains the highest performance, given the similarity of the pre-training task. Remarkably, the LEE approach, employing a distinct training objective, approaches the performance of the Siamese network, showcasing its robustness.

To assess the models' adaptability to entirely new data, we evaluated the English models on the LP and the Finnish models on the LibriSpeech data, as indicated in the last two columns of Table V. Surprisingly, the Speech2vec model, trained on Finnish data and evaluated on English, outperforms its results on LibriSpeech data. This phenomenon suggests that the model generalises more effectively when a large vocabulary is available during training, supported by the considerable vocabulary difference in the LP corpus (over 107 K words appearing only once), as outlined in Table VI. On the other hand, the Audio word2vec model, trained on LP and evaluated on LibriSpeech, demonstrates diminished performance, implying that this model benefits more from a small but clean vocabulary than a larger,



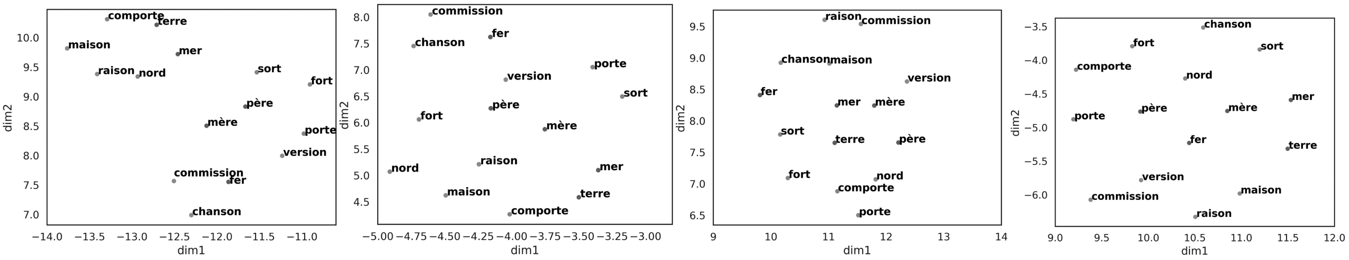
(a) Audio word2vec. (b) Speech2vec. (c) Siamese. (d) LEE.

Fig. 1. Visualisation of the audio embeddings extracted from English LibriSpeech.



(a) Audio word2vec. (b) Speech2vec. (c) Siamese. (d) LEE.

Fig. 2. Visualisation of the audio embeddings extracted from Finnish Lahjoita Puhetta.



(a) Audio word2vec. (b) Speech2vec. (c) Siamese. (d) LEE.

Fig. 3. Visualisation of the audio embeddings extracted from French Common Voice.

noisier one. Despite the observed performance degradation on unseen data, the Siamese and LEE models, though less robust, still achieve superior results.

Overall, these experiments indicate that the explored small-scale speech embedding approaches are not language-dependent and can perform well on unseen data. Furthermore, some approaches demonstrate enhanced performance when trained on a different language, suggesting potential benefits from combining both datasets.

To delve deeper into the ability of the models to distinguish similar and dissimilar sounding words, we visualised their vectors in a 2D space using the UMAP algorithm [66]. Figs. 1, 2, and 3 depict the English, Finnish, and French scatter plots. In the English case, the Siamese and LEE plots reveal more well-defined clusters, consistent with the word discrimination results. Conversely, the Finnish plots exhibit less distinct clusters, possibly due to higher edit distances among Finnish words (e.g., "hetkinen," "semmonen," "sellainen"). Additionally, the accent on the first syllable in the Finnish words contributes to variations in sound, even among words with the same ending. On the French plot, we can observe again that the Siamese and LEE models have better defined clusters than the other models. Despite that, the clustering is still better for the English models. One possible reason for the lower-quality clusters could be due to the considerably smaller amount of training data used to learn the embeddings.

In contrast to the small-scale solutions, the pre-trained transformer models operate on whole utterances. To assess the discriminative capabilities of these models, we conducted an emotion discrimination test, as outlined in Table VII.

The results obtained for the base models show that for English, the Wav2vec2 model significantly outperforms the other variants. Interestingly, for Finnish, the Whisper model produced the highest score, even though it gave the worst results on the Emotion ID task (Table III). For French, the HuBERT model produced the best results, even though it has not seen French data during pre-training.

TABLE VII
ACCURACY SCORES ON THE EMOTION DISCRIMINATION TASK

Embeddings	English	Finnish	French
Base models			
Wav2vec2-B	**70.1**	72.4	75.4
HuBERT-B	66.4	69.6	**89.7**
WavLM-B+	67.2	73.4	81.7
Whisper-small	67.6	**75.3**	85.7
Large models			
Wav2vec2-L	68.1	59.9	83.3
HuBERT-L	63.1	64.5	**87.3**
WavLM-L	**71.0**	**70.7**	**87.3**
Large ASR fine-tuned models			
Wav2vec2-FT	**67.5**	**67.2**	80.2
HuBERT-FT	66.1	**67.2**	**90.5**
WavLM-FT	61.8	62.6	88.1

The bolded values indicate the highest scores achieved for a particular task and model group.

From the large versions, HuBERT demonstrated a performance drop on all the datasets compared to its base counterpart. In Finnish, all three large models showed a decrease in performance, suggesting that the base versions generalise better on a language not seen during pre-training. In French, however, by using the large Wav2vec2 and WavLM versions, we observed improvements.

The extra language-specific data during ASR fine-tuning was beneficial for some while erroneous for the other models. For all the languages, the HuBERT model benefited from the additional data. The Wav2vec2 results improved significantly for Finnish when we used the fine-tuned version but degraded slightly for the other languages. The large improvement in Finnish could be attributed to the additional Finnish data used during the ASR fine-tuning. We observed the highest performance drop for the English and Finnish WavLM models, suggesting a potential discriminatory loss during ASR fine-tuning. Generally, the base models, having significantly fewer parameters, showed comparable results to their larger counterparts, making them a better choice for discriminative tasks.

D. Compressive Representation Results

From the results observed so far, it is evident that the pre-trained transformer models outperform the small-scale solutions, although at a higher computational cost (discussed in the following subsection). Drawing inspiration from the study in [67], we anticipated similar performance outcomes when utilising only the most important dimensions and discarding the rest. To identify these crucial dimensions, we used the Integrated Gradients method.

Fig. 4 depicts sorted absolute attribution values for each dimension on the Intent ID task using the base Wav2vec2 model. Based on the figure, we can observe an exponential drop in the absolution values, indicating that only a small amount of the embedding dimensions encode the relevant task information. We found similar observations for the Gender and Emotion ID tasks.

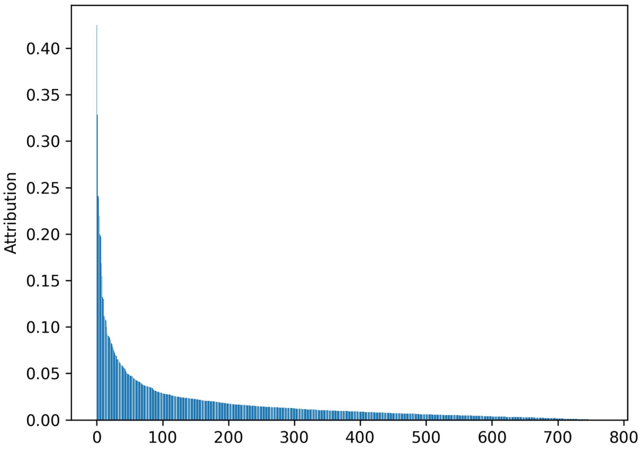


Fig. 4. Absolute attribution values for each dimensions of the base Wav2vec2 model on the Intent ID task.

TABLE VIII
MODELS TRAINED AND EVALUATED ON THE ENGLISH DATASETS WITH 10% CAPACITY

	Gender		Emotion		Intent	
Embeddings	F1	Change	F1	Change	F1	Change
Base models						
Wav2vec2-B	98.2	1.2% ↑	55.9	6.8% ↓	53.3	13.0% ↓
HuBERT-B	97.6	0.4% ↑	58.7	13.8% ↓	58.9	6.5% ↓
WavLM-B+	97.9	1.8% ↑	59.4	11.0% ↓	66.3	5.8% ↓
Whisper-small	98.6	0.2% ↓	54.5	11.4% ↑	30.1	46.1% ↓
Large models						
Wav2vec2-L	98.2	0.1% ↓	56.1	15.8 ↓	61.3	4.4% ↓
HuBERT-L	98.4	0.2% ↑	58.1	9.9% ↓	41.6	34.1% ↓
WavLM-L	97.9	2.4% ↑	60.4	7.0% ↓	35.9	50.4% ↓
Large ASR fine-tuned models						
Wav2vec2-FT	97.9	2.3% ↑	59.2	2.7% ↓	65.3	7.8% ↓
HuBERT-FT	98.0	0.4% ↑	55.7	12.5% ↓	65.1	10.3% ↓
WavLM-FT	98.2	1.0% ↑	61.3	8.3% ↓	60.2	11.4% ↓

Table VIII shows how the pre-trained models perform on the English Gender, Emotion and Intent ID tasks when using only 10% of their capacity. Surprisingly, by reducing the capacity of the models on the Gender ID task, we often observed a performance increase. These findings indicate that for simple tasks, such as Gender ID, only a small number of dimensions encode relevant information, while the others add noise.

On the Emotion ID task, however, we observed a performance drop for all the models when restricting them to 10% of their most important dimensions. An exception here is the Whisper model, which got an 11.4% improvement. The results also suggest that the dimensions of the large and fine-tuned models are better separated, resulting in a smaller performance drop.

On the most challenging Intent ID task, the base models suffer less with reduced capacity. The Whisper, large Wav2vec2 and HuBERT models experienced a large performance drop, suggesting that in those models, for more complex tasks, the relevant information is encoded in more regions. However, by ASR fine-tuning them, the performance drop is less prominent,

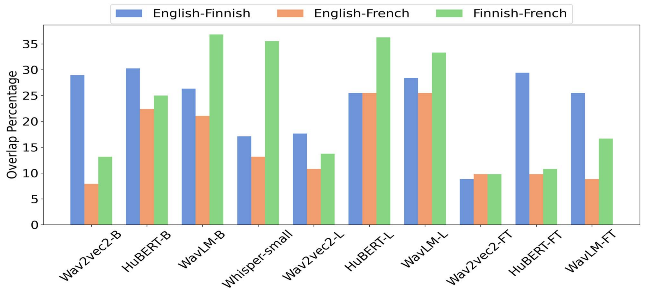


Fig. 5. Dimension overlap across languages for Gender ID.

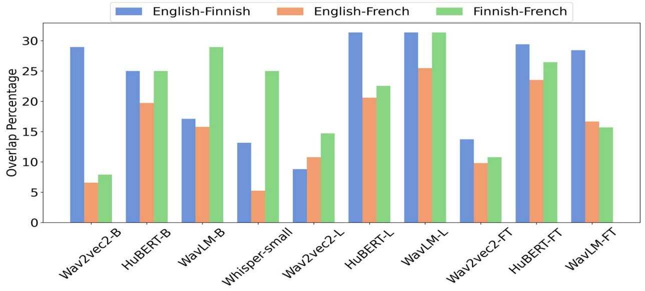


Fig. 6. Dimension overlap across languages for Intent ID.

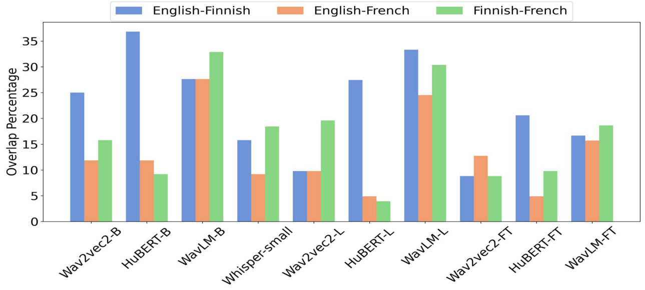


Fig. 7. Dimension overlap across languages for Intent/Topic/Request ID.

suggesting that the ASR tasks help with separating the dimensions.

Overall, most models perform competitively, or even better when restricted to only 10% of their most important dimensions, making it a viable option when the computational resources are constrained.

In the next experiment, we delved into the relationship between the key dimensions across the languages by calculating the dimension overlap. When comparing the dimension overlap across the languages (Figs. 5, 6, 7), we found that for Gender ID (Fig. 5), the base WavLM model has the highest overlap of 36.8% between Finnish and French, closely followed by the large HuBERT and Whisper models. Interestingly, despite their linguistic proximity, English and French showed the lowest overlap.

On the Emotion ID task, depicted in Fig. 6, we found the agreement between the languages to be even smaller. We observed the highest agreement of 31.4% between English and Finnish using the large HuBERT and WavLM and between Finnish and French using the large WavLM model. Notably, for this task, the agreement between English and Finnish tends to be higher, which could indicate a similarity in the way the emotions are expressed in those languages.

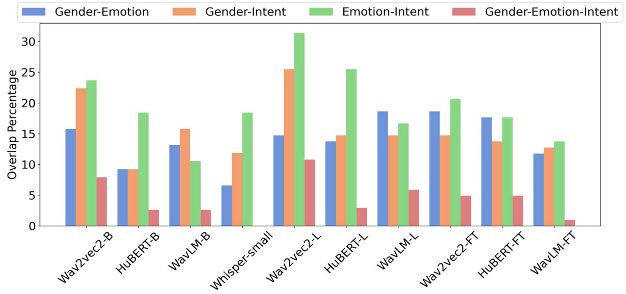


Fig. 8. Dimension overlap across tasks on the English datasets.

For the Intent, Topic, and Request ID tasks (Fig. 7), the base HuBERT model demonstrated the highest agreement at 36.8% between English and Finnish, but it dropped significantly between English and French (11.8%) and Finnish and French (9.2%). Conversely, all three WavLM models showed relatively high agreements across languages, suggesting that for this model, similar regions are responsible for the task, potentially learning more universal features, regardless of the language.

By examining the overlap between the 10% most important dimensions across languages, we found that those dimensions often differ. These findings suggest that the crucial regions for each task might depend on the language (or dataset) used during pre-training. We believe this to be the case due to the higher agreement for the models that do not have a multilingual or language-specific variant, such as the WavLM base. We observed a similar behaviour for the large HuBERT version, which does not have a Finnish variant, resulting in a high dimension overlap. Additionally, the WavLM model showed high overlap on all three tasks, suggesting that it has learned language-agnostic representations.

Besides investigating the dimension overlap across languages, we investigated whether some dimensions are shared across tasks. To achieve that, we calculated the dimension overlap between Gender, Emotion and Intent ID tasks for each model, as presented in Fig. 8.

The results revealed that the highest agreement is between Emotion and Intent ID for all the models except the WavLM base and large variants. These findings hint that the dimensions responsible for the prosodic information play an important role in semantic understanding.

Another interesting finding is that in many cases, the highest overlap is between Gender and Intent ID, compared to Gender and Emotion ID tasks, even though Gender and Intent ID have much less in common. These findings point to potential spurious correlations that the models have learned during the pre-training phase.

When comparing the agreement across all three tasks, we found that the Wav2vec2 variants have the highest dimension overlap across all three tasks, with the base version having the highest of 7.9%.

TABLE IX
NUMBER OF PARAMETERS (IN THE EMBEDDING MODELS) AND INFERENCE TIME ON THE IEMOCAP EMOTION ID TEST SET, WITH A BATCH SIZE OF 1

Embeddings	**Parameters**	**Time(s)**		
Small-scale models		Full	Short	Long
Audio word2vec	270K	2.3	/	/
Speech2vec	287K	2.3	/	/
Siamese	146K	2.3	/	/
LEE	910K	2.3	/	/
Base models				
Wav2vec2-B	95M	101	0.8	2.3
HuBERT-B	95M	100	0.8	2.3
WavLM-B plus	95M	103	0.8	2.4
Whisper-small	244M	277	3.0	3.0
Large models				
Wav2vec2-L	317M	114	0.9	3.1
HuBERT-L	317M	115	1.0	3.2
WavLM-L	317M	119	1.0	3.4
Large ASR fine-tuned models				
Wav2vec2-FT	317M	114	1.0	3.2
HuBERT-FT	317M	115	0.9	3.2
WavLM-FT	317M	119	1.0	3.4

The short samples are under 3 seconds and the long ones are over 15. Both short and long subsets have 15 samples.

E. Inference Time

In the final series of experiments, we assessed the inference time for each model on the IEMOCAP test set, with the results outlined in Table IX. We conducted these experiments with a batch size of 1 on a consumer-grade NVIDIA GeForce RTX 2080 GPU.

The small-scale models have identical inference times, as expected, given their utilisation of a same-size encoder for extracting speech embeddings. It is important to emphasise that the pre-computed MFCC features, used as input to the encoders, are not factored into the inference time.

Examining the base versions of the large pre-trained models on the full test set, the Whisper-small model measures the longest inference time at 277 seconds, while the remaining models demonstrate similar performance - a predictable outcome considering their analogous architectures and parameters. As anticipated, the large fine-tuned models, with their substantially increased parameters, come with higher inference times compared to their base counterparts; however, the difference is not big.

We can observe a small difference between the large and base versions on the subset consisting of short samples (smaller than 3 seconds). On utterances longer than 15 seconds, however, the difference between the base and large models is more substantial. Interestingly, the Whisper model performs equally with both short and long segments.

In summary, the small-scale models, characterised by a considerably lower number of parameters, exhibit notably faster inference times. Nevertheless, this efficiency comes at the expense of reduced performance, as evidenced in the intrinsic and extrinsic assessments. Consequently, selecting an appropriate model for a given task requires a careful trade-off consideration between inference time and performance.

IX. CONCLUSION

Due to the abundance of speech embedding extraction methods, choosing the right one is difficult for industry and research applications. To ease the process of selecting the appropriate model, in this study, we thoroughly evaluated small and large-scale speech embedding approaches in intrinsic and extrinsic ways in English, Finnish, and French languages.

The discriminative intrinsic experiments revealed that the compact, small-scale Siamese and LEE models perform exceptionally well, although suffering performance degradation when evaluated on unseen vocabulary. Emotion discrimination highlighted the effectiveness of the base self-supervised models despite having fewer parameters. For tasks that require modelling the prosodic information, the base self-supervised models performed comparably to the larger ones on Gender ID, making them preferable. MFCC features, though less computationally demanding, were slightly worse. The large HuBERT model excelled on Finnish Emotion ID without language-specific pre-training, showing strong language-agnostic capabilities. Semantic assessments via Intent, Topic, and Request ID tasks revealed that the base large-scale self-supervised models often matched or exceeded their larger counterparts. The WavLM model, pre-trained on English, performed best in Finnish, demonstrating its adaptability to new languages.

We made multiple discoveries by conducting experiments with the most crucial dimensions for each task. For simple tasks like Gender ID, limiting the models to their top 10% most important dimensions improved the performance, suggesting that many dimensions add noise. On Emotion ID, the large and fine-tuned models showed better dimension separation, resulting in less performance drop. However, for the more complex Intent ID task, the base models had smaller performance drops, while the large Wav2vec2, HuBERT, and Whisper variants showed significant degradation, indicating they encode relevant information across multiple regions. ASR fine-tuning mitigated the performance drop, implying that ASR aids in dimension separation. The dimension overlap analysis revealed that crucial dimensions often differ between languages, depending on the pre-training data. By measuring the overlap between the tasks we showed that there is a higher overlap between Gender and Intent ID than between Gender and Emotion ID, despite the former being less related, suggesting possible spurious correlations learned during pre-training.

APPENDIX

Figs. 9, 10, and 11 show the layer analysis conducted for the Gender, Emotion, and Intent ID tasks in English. The experiments are conducted on a subset of the training data using the base, large, and ASR fine-tuned versions. We determined the best layer for each model based on the highest F1 score.

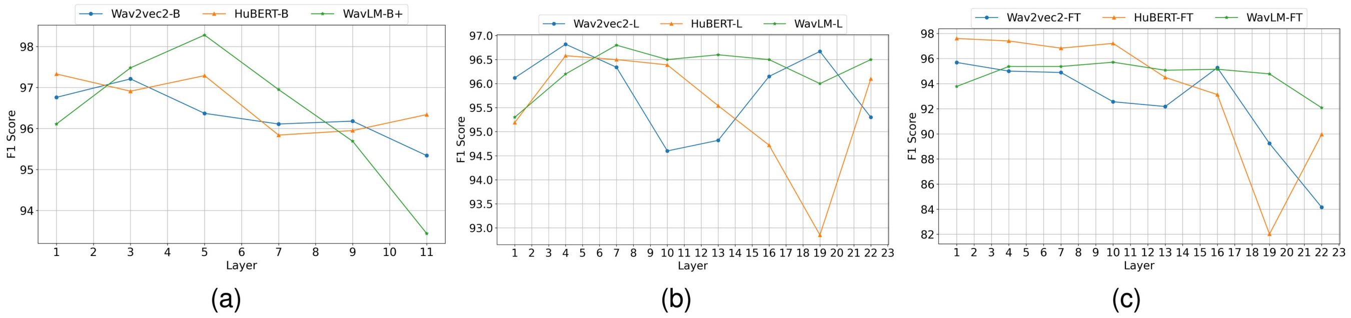


Fig. 9. F1 scores obtained using individual layers on the Gender ID task in English for the base (a), large (b) and ASR fine-tuned (c) versions.

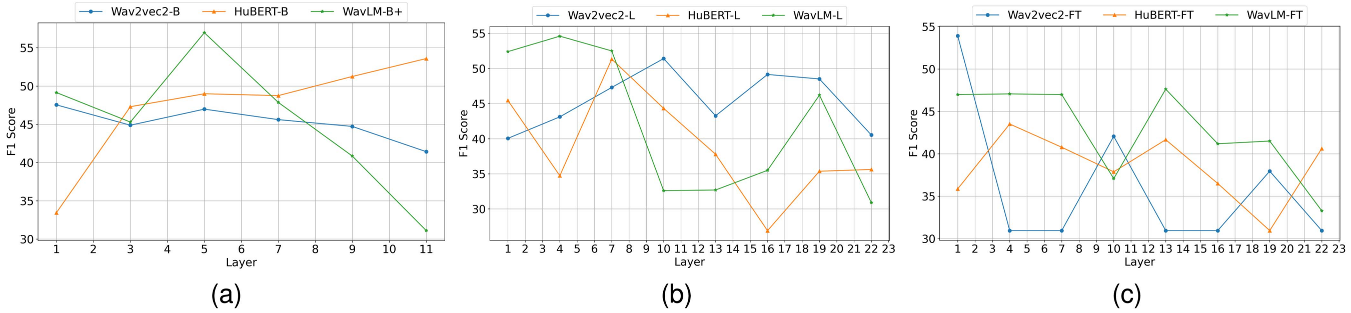


Fig. 10. F1 scores obtained using individual layers on the Emotion ID task in English for the base (a), large (b) and ASR fine-tuned (c) versions.

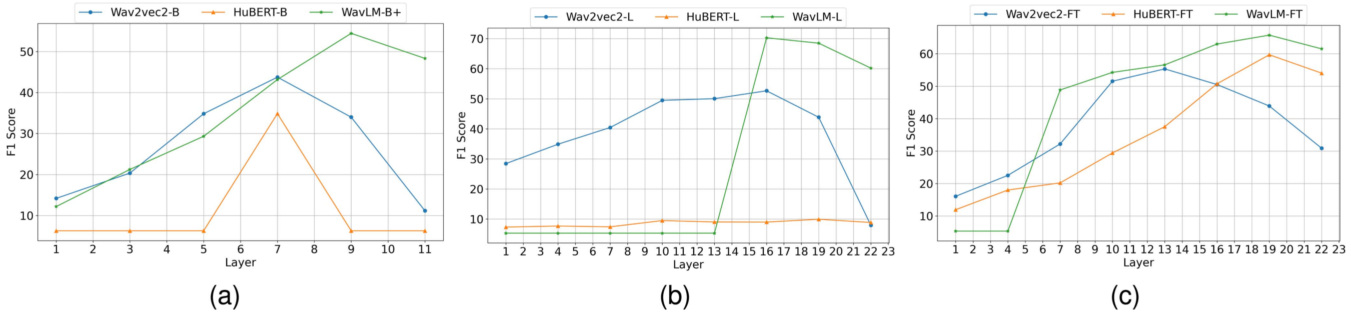


Fig. 11. F1 scores obtained using individual layers on the Intent ID task in English for the base (a), large (b) and ASR fine-tuned (c) versions.

REFERENCES

[1] D. Serdyuk, Y. Wang, C. Fuegen, A. Kumar, B. Liu, and Y. Bengio, "Towards end-to-end spoken language understanding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5754–5758.

[2] M. Radfar, A. Mouchtaris, and S. Kunzmann, "End-to-end neural transformer based spoken language understanding," in *Proc. Interspeech*, 2020, pp. 866–870.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.

[4] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, 2017. [Online]. Available: https://aclanthology.org/Q17-1010

[5] M. Ulčar et al., "High quality ELMo embeddings for seven less-resourced languages," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 4731–4738. [Online]. Available: https://aclanthology.org/2020.lrec-1.582

[6] S. Papay, S. Padó, and N. T. Vu, "Addressing low-resource scenarios with character-aware embeddings," in *Proc. 2nd Workshop Subword/Character LEvel Models*, 2018, pp. 32–37.

[7] B. Milde and C. Biemann, "Unspeech: Unsupervised speech context embeddings," in *Proc. Interspeech*, 2018, pp. 2693–2697.

[8] Y.-A. Chung, C.-C. Wu, C.-H. Shen, H.-Y. Lee, and L.-S. Lee, "Audio Word2Vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder," in *Proc. Interspeech*, 2016, pp. 765–769.

[9] Y.-A. Chung and J. Glass, "Speech2Vec: A sequence-to-sequence framework for learning word embeddings from speech," in *Proc. Interspeech*, 2018, pp. 811–815.

[10] B. M. Abdullah, M. Mosbach, I. Zaitova, B. Möbius, and D. Klakow, "Do acoustic word embeddings capture phonological similarity? An empirical study," in *Proc. Interspeech*, 2021, pp. 4194–4198.

[11] A. Nandan and J. Vepa, "Language agnostic speech embeddings for emotion classification," in *Proc. ICML Workshop Self-Supervision Audio Speech*, 2020. [Online]. Available: https://openreview.net/forum?id=jaXJWbbBvG_

[12] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a 'siamese' time delay neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1993, vol. 6, pp. 737–744.

[13] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[14] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, " BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[16] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel, "Strategies for automatic segmentation of audio data," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Proc.*, 2000, vol. 3, pp. 1423–1426.

[17] Y.-C. Chen, S.-F. Huang, C.-H. Shen, H.-y. Lee, and L.-s. Lee, "Phonetic-and-semantic embedding of spoken words with applications in spoken content retrieval," in *Proc. IEEE Spoken Lang. Technol. Workshop*,2018, pp. 941–948.

[18] A. Haque, M. Guo, P. Verma, and L. Fei-Fei, "Audio-linguistic embeddings for spoken sentences," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 7355–7359.

[19] M. A. Carlin, S. Thomas, A. Jansen, and H. Hermansky, "Rapid evaluation of speech representations for spoken term discovery," in *Proc. 12th Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 821–824.

[20] K. Chen, B. Liang, X. Ma, and M. Gu, "Learning audio embeddings with user listening data for content-based music recommendation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 3015–3019.

[21] H. Kamper, W. Wang, and K. Livescu, "Deep convolutional acoustic word embeddings using word-pair side information," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2016, pp. 4950–4954.

[22] S. Settle and K. Livescu, "Discriminative acoustic word embeddings: Recurrent neural network-based approaches," in *Proc. IEEE Spoken Lang. Technol. Workshop*. 2016, pp. 503–510.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] W. He, W. Wang, and K. Livescu, "Multi-view recurrent neural acoustic word embeddings," in *Proc. Int. Conf. Learn. Representations*, 2017.

[25] A. L. Maas, S. D. Miller, T. M. O'neil, A. Y. Ng, and P. Nguyen, "Word-level acoustic modeling with convolutional vector regression," in *Proc. ICML Workshop Representation Learn.*, 2012.

[26] G. Chen, C. Parada, and T. N. Sainath, "Query-by-example keyword spotting using long short-term memory networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2015, pp. 5236–5240.

[27] S. Mozziconacci, "Pitch variations and emotions in speech," in *Proc. XIIIth Int. Congr. Phonetic Sci.*, 1995, vol. 1, pp. 178–181.

[28] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, *arXiv:1803.01271*.

[29] T. Brown et al., "Language models are few-shot learners," *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 1877–1901.

[30] S. Hershey et al., "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 131–135.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.

[32] G. Cerutti, R. Prasad, A. Brutti, and E. Farella, "Neural network distillation on IoT platforms for sound event detection," in *Proc. Interspeech*, 2019, pp. 3609–3613.

[33] H. Xie and T. Virtanen, "Zero-shot audio classification via semantic embeddings," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1233–1242, 2021.

[34] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 12449–12460.

[35] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Proc. Interspeech*, 2021, pp. 3400–3404.

[36] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT : Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[37] S. Chen et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.

[38] R. Algayres, M. S. Zaiem, B. Sagot, and E. Dupoux, "Evaluating the reliability of acoustic speech embeddings," in *Proc. Interspeech*, 2020, pp. 4621–4625.

[39] T. Ashihara et al., "SpeechGLUE: How well can self-supervised speech models capture linguistic knowledge?," in *Proc. Interspeech*, 2023, pp. 2888–2892.

[40] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. EMNLP Workshop Blackbox NLP: Analyzing Interpreting Neural Netw. NLP,* 2018, pp. 353–355. [Online]. Available: https://aclanthology.org/W18-5446

[41] S. wen Yang et al., "SUPERB: Speech processing universal PERformance benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

[42] J. Shi et al., "ML-SUPERB: Multilingual speech universal PERformance benchmark," in *Proc. Interspeech*, 2023, pp. 884–888.

[43] A. Mohamed et al., "Self-supervised speech representation learning: A review," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1179–1210, Oct. 2022.

[44] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2015, pp. 5206–5210.

[45] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, pp. 335–359, 2008.

[46] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, "SLURP: A spoken language understanding resource package," in *Proc.Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 7252–7262. [Online]. Available: https://aclanthology.org/2020.emnlp-main.588

[47] A. Moisio et al., "Lahjoita puhetta: A large-scale corpus of spoken Finnish with some benchmarks," *Lang. Resour. Eval.*, vol. 57, pp. 1295–1332, 2022.

[48] M. Airas and P. Alku, "Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalised amplitude quotient," *Phonetica*, vol. 63, no. 1, pp. 26–46, 2006.

[49] E. Vaaras, M. Airaksinen, and O. Räsänen, "Analysis of self-supervised learning and dimensionality reduction methods in clustering-based active learning for speech emotion recognition," in *Proc. Interspeech*, 2022, pp. 1143–1147.

[50] R. Ardila et al., "Common voice: A massively-multilingual speech corpus," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 4218–4222. [Online]. Available: https://aclanthology.org/2020.lrec-1.520

[51] P. Gournay, O. Lahaie, and R. Lefebvre, "A Canadian french emotional speech dataset," in *Proc. 9th ACM Multimedia Syst. Conf.*, 2018, pp. 399–402, doi: 10.1145/3204949.3208121.

[52] N. Lackovic, C. Montacié, G. Lalande, and M.-J. Caraty, "Prediction of user request and complaint in spoken customer-agent conversations," 2022, *arXiv:2208.10249*.

[53] B. W. Schuller et al., "The ACM multimedia 2023 computational paralinguistics challenge: Emotion share & requests," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 9635–9639.

[54] F. A. Gers and J. Schmidhuber, "Recurrent nets that time and count," in *Proc. IEEE-INNS-ENNS Int. Joint Conf. Neural Netw. . Neural Comput.: New Challenges Perspectives New Millennium*, 2000, vol. 3 pp. 189–194.

[55] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. 12, pp. 3371–3408, 2010.

[56] S. Subramanian, A. Trischler, Y. Bengio, and C.J. Pal, "Learning general purpose distributed sentence representations via large scale multi-task learning," in *Proc. Int. Conf. Learn. Representations*, 2018.

[57] D. Porjazovski, T. Grósz, and M. Kurimo, "Topic identification for spontaneous speech: Enriching audio features with embedded linguistic information," in *Proc. 31st Eur. Signal Process. Conf.*, 2023, pp. 396–400.

[58] Z. Chi et al., "XLM-E: Cross-lingual language model pre-training via ELECTRA," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 6170–6182. [Online]. Available: https://aclanthology.org/2022.acl-long.427

[59] T. Grósz, D. Porjazovski, Y. Getman, S. Kadiri, and M. Kurimo, "Wav2vec2-based paralinguistic systems to recognise vocalised emotions and stuttering," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 7026–7029.

[60] N. Naderi and B. Nasersharif, "Cross corpus speech emotion recognition using transfer learning and attention-based fusion of Wav2Vec2 and prosody features," *Knowl.-Based Syst.*, vol. 277, 2023, Art. no. 110814.

[61] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.* 2017, pp. 3319–3328.

[62] D. Povey et al., "The Kaldi speech recognition toolkit," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2011.

[63] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2021, pp. 914–921.

[64] Y. Wang, A. Boumadane, and A. Heba, “A fine-tuned wav2vec 2.0/Hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding,” 2021. *arXiv:2111.02735*.

[65] A. Pasad, B. Shi, and K. Livescu, “Comparative layer-wise analysis of self-supervised speech models,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5.

[66] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform manifold approximation and projection for dimension reduction,” 2018, *arXiv:1802.03426*.

[67] T. Grósz, A. Virkkunen, D. Porjazovski, and M. Kurimo, “Discovering relevant sub-spaces of BERT, wav2vec 2.0, ELECTRA and ViT embeddings for humor and mimicked emotion recognition with integrated gradients,” in *Proc. 4th Multimodal Sentiment Anal. Challenge Workshop: Mimicked Emotions Humour Personalisation*, 2023, pp. 27–34.

Dejan Porjazovski received the master’s degree in machine learning, data science, and artificial intelligence from Aalto Univeristy, Espoo, Finland, in 2020. After his master’s degree, he started his Doctoral studies with Aalto University with the Automatic Speech Recognition group, led by Professor Mikko Kurimo. His research focuses on spoken language understanding for low-resource languages.

Tamás Grósz (Member, IEEE) received the Ph.D. degree in speech recognition from the University of Szeged, Szeged, Hungary, in 2018. From 2017 and 2018, he was as an Assistant Research Fellow with the Hungarian Academy of Sciences’ Research Group on Artificial Intelligence. From 2018 to 2019, he was a Senior Lecturer with the Department of Computer Algorithms and Artificial Intelligence, University of Szeged. He is currently a Research Fellow with the Department of Information and Communications Engineering, Aalto University, Espoo, Finland. His research interests include automatic speech recognition, deep learning, computational paralinguistics, and explainable AI.

Mikko Kurimo (Senior Member, IEEE) received the D.Sc.Tech. degree in 1997. He is currently a Full Professor of speech and language processing and Head of the Speech Recognition Group, Aalto University, Espoo, Finland. He has supervised 18 doctoral theses and 76 masters thesis and coordinated several large national (Academy of Finland and Business Finland) and international (EC and Nordforsk) research projects. He has authored or coauthored more than 240 peer reviewed international conference and journal publications. His research interests include machine learning in speech and language technology. His groups research results have been widely utilized via their open source tools and the achievements include, such as, winning the MGB-3 challenge for building low-resourced dialect ASR system and Compare 2022 stuttering and non-verbal vocalizations challenges and Compare 2023 spoken emotions challenge.