Neural Computing and Applications
https://doi.org/10.1007/s00521-024-09435-1

ORIGINAL ARTICLE

Exploring emergent syllables in end-to-end automatic speech recognizers through model explainability technique

Vincenzo Norman Vitale[1] · Francesco Cutugno[1] · Antonio Origlia[1] · Gianpaolo Coro[2]

Received: 22 May 2023 / Accepted: 15 January 2024

Abstract
Automatic speech recognition systems based on end-to-end models (E2E-ASRs) can achieve comparable performance to conventional ASR systems while reproducing all their essential parts automatically, from speech units to the language model. However, they hide the underlying perceptual processes modelled, if any, and they have lower adaptability to multiple application contexts, and, furthermore, they require powerful hardware and an extensive amount of training data. Model-explainability techniques can explore the internal dynamics of these ASR systems and possibly understand and explain the processes conducting to their decisions and outputs. Understanding these processes can help enhance ASR performance and reduce the required training data and hardware significantly. In this paper, we probe the internal dynamics of three E2E-ASRs pre-trained for English by building an acoustic-syllable boundary detector for Italian and Spanish based on the E2E-ASRs' internal encoding layer outputs. We demonstrate that the shallower E2E-ASR layers spontaneously form a rhythmic component correlated with prominent syllables, central in human speech processing. This finding highlights a parallel between the analysed E2E-ASRs and human speech recognition. Our results contribute to the body of knowledge by providing a human-explainable insight into behaviours encoded in popular E2E-ASR systems.

Keywords Automatic speech recognition · Deep learning · End-to-end models · Long short-term memory model · Conformers · Transformers · Psycho-acoustics · Syllables

1 Introduction

Deep-learning architectures are becoming increasingly complex, and their required training datasets are increasingly extensive [1]. Automatic speech recognition (ASR) systems based on deep-learning models (DL-ASRs) can reach very high performance on controlled speech and have opened the way to a vast range of voice-activated applications [2–4]. However, these powerful models only partially disclose how they achieve speech recognition. They leave open questions such as (i) which underlying rhythmic processes are being modelled, if any, and (ii) whether their unexplained, *brute force* modelling approach could be substituted by a psycho-acoustic-inspired and less hardware-demanding approach. These aspects are essential to enhance ASR performance and efficiency. Psycho-acoustic-inspired ASRs can indeed achieve comparable performance to DL-ASRs with much less (even of orders of magnitude) training material and computational time and energy expense [5]. Moreover, human speech recognition's effectiveness strongly depends on the supra-segmental processing of speech (including rhythm). Human speech processing has indeed been explored from multiple perspectives in recent decades. In particular, the role of rhythmic scan has attracted interest as a way to describe attentional patterns supporting acoustic information weighting in real time [6, 7]. The concept of *syllable* has

✉ Gianpaolo Coro
gianpaolo.coro@isti.cnr.it

Vincenzo Norman Vitale
vincenzonorman.vitale@unina.it

Francesco Cutugno
cutugno@unina.it

Antonio Origlia
antonio.origlia@unina.it

[1] UrbanECO, Università degli Studi di Napoli Federico II, Corso Umberto I, 40, 80138 Naples, Italy

[2] Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo", CNR, Via Moruzzi, 1, 56124 Pisa, Italy

Published online: 13 February 2024

been extensively used to describe acoustic correlates of rhythmic patterns. From a phonetic point of view, a *syllable* has been defined [8] as "a continuous voiced segment of speech organised around one local loudness peak, and possibly preceded and/or followed by voiceless segments". In a more acoustic-oriented definition [9, p. 70] that considers co-articulation dynamics, the syllable consists of "a centre which has little or no obstruction to airflow and which sounds comparatively loud; before and after that centre [...] there will be a greater obstruction to airflow and/or less loud sound". Consequently, a *syllable* has been described as a 100–250-ms signal segment constructed around a high energy peak (*nucleus*), possibly preceded by an increasing energy slope (*onset*) and followed by a tail of decreasing energy (*coda*). Syllables have been used as the basis for theories describing how speech processing has developed in human beings over time [10], and several studies have highlighted their crucial importance in speech perception and recognition [11–16]. Syllables can indeed be perceived even if they are reduced or not actually uttered [17–19]. However, syllabic speech units pose difficulties in linguistics and psycho-acoustics, e.g. different language experts sometimes disagree on positioning their boundaries [20, 21].

A high-performance DL-ASR might internally reproduce linguistic information at different levels of abstraction; in particular, we argue that one of these possible representations could be related to acoustic syllables 'spontaneously', in a way that could be sufficient to teach another automatic system to recognise these syllables simply using the internal vector embeddings as features. These representations could be found in the deep-learning model's encoding layers and would be automatically formed, while the model learns to recognise speech units and language [22]. End-to-end (E2E) models belong to this class of deep-learning models (Sect. 2). They automatically model all speech recognition mechanisms, from base speech units to the language model. Through model-explainability techniques, it is possible to verify if the specific internal dynamics of these models resemble those of human speech perception. If these dynamics exist, they can be explored, understood, and explicitly re-embedded in the ASR model. This approach can lead to significant technological breakthroughs because a psycho-acoustics-inspired ASR model would require much less training data and simpler hardware to achieve a high performance [5]. This research would thus support a more computationally accessible artificial intelligence, a relevant problem in modern technology [23, 24] (Sect. 2).

In the present paper, we use the *syllable* as the central acoustic unit of an investigation of end-to-end ASR models (E2E-ASRs). We analyse the internal learning processes of a single E2E-ASR architecture with three different sizes of internal encoding and decoding modules. We demonstrate that these models automatically developed in their shallower layers a model of rhythmic patterns related to syllables. These patterns resembled syllabic-scale human speech perception processes that past linguistic and psycho-acoustic studies have also described [12, 13, 25–32]. To this aim, we built automatic syllable boundary detectors working on the vectors extracted from the internal ASR models' encoding layers. These detectors allowed us to identify the layers where syllable-related information was formed and calculate rhythmic and intensity properties.

This paper is organised as follows: Sect. 2 reports background and related work on deep-learning-based ASRs and their explainability. Section 3 describes the end-to-end ASRs used in our experiment, the syllable boundary detectors we built, and the data we used for the evaluation. Section 4 reports the performance of our syllable boundary detectors and supports the demonstration that the inner layers of the used ASRs contain syllable-related information. Finally, Sect. 5 draws the conclusions.

2 Background and related works

An E2E-ASR automatically transforms a sequence of input acoustic feature vectors (possibly raw samples) from an audio signal into a sequence of graphemes or words representing the audio transcription [3]. Conventional ASR systems usually train acoustic, pronunciation, and language models separately and require specific modelling and training of these parts. E2E-ASRs overcome the difficulties and cost-ineffectiveness of the data preparation and modelling phases of conventional systems, by committing the model to learn all parts automatically. E2E-ASRs can perform comparable to conventional systems but require far more training data [5].

Despite the great interest by academics and industry in E2E-ASRs, their usage in production environments encountered obstacles due to practical issues like insufficient client streaming capabilities, high latency, and low multi-application-context adaptability [33]. Moreover, having all information hidden in a complex deep learning model limits understanding the model's internal dynamics and the confidence in using an E2E-ASR for commercial or industrial applications. Additionally, the continuous increment in the models' size (i.e. the number of layers and parameters) limits their executions to powerful machines only—usually residing in cloud computing infrastructures—rather than to the local users' computers or small embeddable devices. The required data volume and computational capacity to handle model training from scratch increase with model complexity and size so fast that significant investments are necessary to train even one model.

Today, this trend allows only a few institutions and companies to develop state-of-the-art E2E-ASRs.

This section describes the deep-learning models typically used in E2E-ASRs (Sect. 2.1). Then, it provides an overview of the methodologies used to explain the internal information representation formed in these models (Sect. 2.2).

2.1 End-to-end automatic speech recognition models

Using E2E models has been a turning point in automatic speech recognition [33]. E2E models enabled the possibility of merging acoustic and language modelling into one system whose task was to convert an input vector sequence into another (Fig. 1). Today, E2E-ASRs are frequently based on Transformer deep-learning architectures [22, 34]. A Transformer processes sequences of acoustic data vectors and automatically models the information in the vector sequence as a whole and the vectors' inter-dependencies. As a result, it automatically infers speech units and the language model. Transformers commonly use an encoder–decoder architecture: The encoder forms an internal representation of the raw input (e.g. speech units-related data) that also contains information on inter-vector relations. A sequence of encoding layers is usually adopted to create more and more abstract data representations. The decoder translates the encoded data into an output vector sequence. A sequence of decoding layers can be used to refine the output sequence iteratively. The final output of a Transformer-based ASR model is the phonetic or word transcription of the audio input. Transformer-based ASRs require a far more extensive training set than conventional ASR systems to achieve comparable performance [35–37].

A Transformer ASR model typically includes a *self-attention* module in its architecture [38–40]. Self-attention estimates the influence of all preceding and subsequent input data vectors when processing each single data vector. This mechanism was introduced to mimic human cognitive attention because it relates one data vector in the input sequence to all its contextual data vectors. Through *reverse engineering* or deep-layer *probing*, it is possible to analyse the outputs of each encoding, decoding, and self-attention layer to understand whether these reflect perception-related processes [41–44].

The Transformer encoder can be implemented as a sequence of ' Conformer' blocks [45] (Fig. 2), each combining a four Feed-Forward Artificial Neural Network sequence with a final normalisation layer. The name 'Conformer' is commonly used to indicate a Transformer with this encoding method. The Transformer decoder can be substituted by (or combined with) a Connectionist Temporal Classification (CTC) model [35] or a Recurrent Neural Network Transducer (RNN-T) [46]. CTC is a non-auto-regressive speech transcription technique which collapses consecutive, all-equal transcription labels (character, word piece, etc.) to one label unless a special label separates these. The result is a sequence of labels shorter or equal to the input vector sequence length. The CTC is one of the most diffused decoding techniques. As non-auto-regressive, it is also considered computationally effective because it requires less time and resources for the training and inference phases. Conversely, the RNN-T (also named *Transducer*) is an auto-regressive speech transcription technique which overcomes CTC's limitations, i.e. non-auto-regressive and limited label sequence length. An RNN-T (also named *Transducer*) is a speech transcription technique which can produce label-transcription sequences

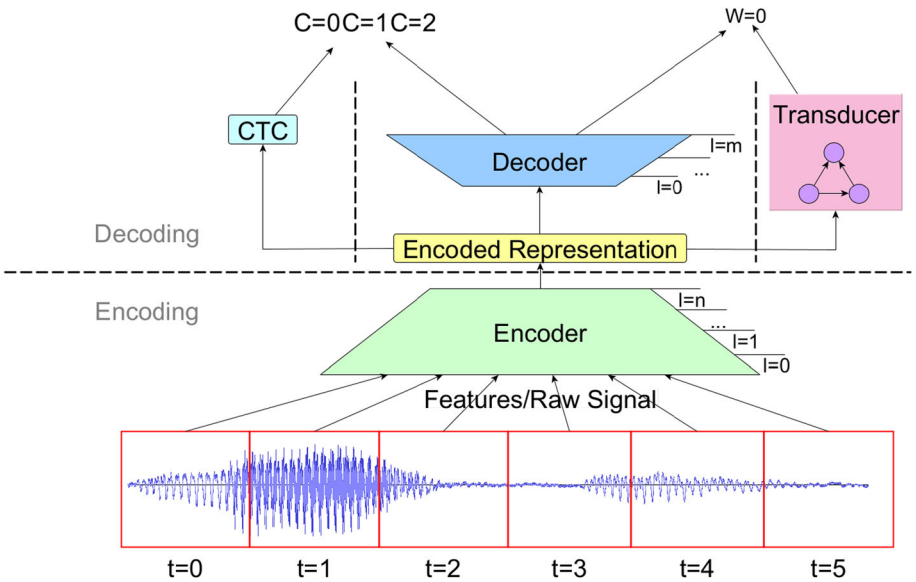

Fig. 1 Example architectural schema of an end-to-end automatic speech recognition model

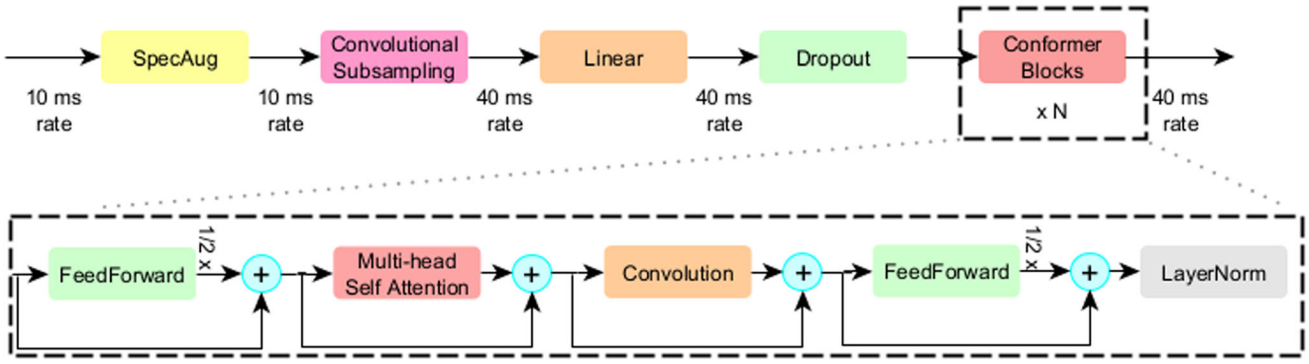

Fig. 2 Typical architectural schema of a Conformer block within a generic deep-learning model's encoder

longer than the input vector sequence and models long-term transcription elements' inter-dependency. A Transducer typically comprises two sub-decoding modules: one that forecasts the next transcription label based on the previous transcriptions (prediction network) and the other that combines the encoder and prediction-network outputs to produce a new transcription label (joiner network). These features improve transcription speed and performance with respect to CTC at the expense of more training and computational resources required [46].

2.2 Model explainability

Among the drawbacks of modern deep learning systems, the most frequently cited are the low accessibility of sufficient training corpora, the high demand for computational resources, and their poor interpretability (i.e. the explanation, understanding, and trust of the decisions and outputs) [47–57]. DL-ASRs are not exempt from these issues [58, 59]. However, the interpretation of the internal model dynamics and overall 'behaviour' can be studied through model-output backtracking or simulated via explainable methods [60–65]. Alternatively, 'probes' can be installed on the encoding (and/or decoding) layers at different 'depths' of the layer sequence [66, 67]. The probes allow observing and then analysing the vectors produced by the layers (*emissions* or emission vectors) to classify some phenomenon and consequently characterise the information contained in these vectors [68]. For example, in computer vision, model probing allows associating focus areas or abstraction patterns to specific deep neural network layers [69].

Probing is often used to interpret specific E2E-ASRs' layers [66, 67]. For example, the DeepSpeech2 E2E-ASR [70] was probed to study the differences between the models trained in English and Arabic [71]. The probed emissions were classified through a Feed-Forward neural network to evidence that the two models were learning specific linguistic characteristics related to articulation manner and place. Other probing studies have analysed, through classification and measurements, how accent influences the DeepSpeech2 performance [72, 73]. These studies also evidenced how the contextual phonetic information contained in the emissions influenced the classification tasks. Probing has also been used to investigate the multi-temporal modelling of phonetic information in the Wav2Vec2.0 E2E-ASR [74, 75]. A recent study [76] has proposed an in-depth analysis of the layer-wise encoded information of the pre-trained Wav2Vec2.0 large and small-sized models. The study evidenced the presence of phone-level and word-level information at layers 11–12/17–18 (phone) and layers 11–19 (word) for the large-sized model. Some studies have also proposed a spectrogram-like representation of emissions that could be used for speaker identification and speech synthesis [77].

Probing studies have seldom analysed the possible multi-scale (e.g. phonetic, syllabic, word), supra-segmental (e.g. rhythm, pitch), modelling occurring in E2E-ASRs, e.g. the presence of syllabic-scale or rhythmic components also existing in human speech recognition [78]. The existence of these components would indicate the presence of interpretable information in the emissions, which could adequately be re-embedded in the ASR model to improve performance while decreasing computational complexity [5]. For example, syllabic information can be primary in data pre-processing for efficiently selecting informative data from large corpora that would improve ASR performance and make it comparable to a system using a much larger amount of data [79]. When used within a maximum-entropy principle for acoustic feature selection and uncertainty quantification [80], syllabic information can help choose training utterances contributing to homogenising information distribution across speech units [81, 82]. Moreover, syllabic spectral analysis can reveal syllabic structural changes related to language evolution and anthropological dynamics [83, 84]. Moreover, syllables are often central speech units in the design of ASRs targeting

under-resourced languages or limited-vocabulary applications [85–90].

3 Materials and method

This section describes our probing of three state-of-the-art Transformer-based ASR models. Rather than searching for word or phonetic scale representations forming within the Transformer encoding layers, we focused on syllable-scale representations. We investigated whether the emissions of the three analysed Transformers were valuable for building a high-performance acoustic-syllable boundary detector. Moreover, we explored whether specific Transformer layers formed syllable-related representations.

The present section is organised as follows: Sect. 3.1 describes the base Transformer ASR models analysed. Section 3.2 describes the syllable boundary detectors we built for the probing task. Finally, Sect. 3.3 describes the data used for training and testing the syllable boundary detectors.

3.1 Transformer ASR models

The Transformer ASR model architecture we probed was a Conformer model from the Nvidia NeMo Automatic Speech Recognition toolkit [91]. Nvidia distributes three pre-trained versions of this ASR model with different 'sizes'—corresponding to different Conformer block sequence lengths—based on the NeMo ASRSET-2.0 open-source corpus in English (Table 1). The NeMo toolkit and the pre-trained models aim to provide academic and industrial researchers with state-of-the-art tools to build conversational agents.

The used Conformers contain a Conformer block sequence in its encoder module (encoding layers). The decoder uses a Transducer for word-based decoding or, alternatively, a CTC model for character-based decoding. For the present study, we used word-based decoding because we addressed the detection of speech units with a larger scale than the phonetic scale (which roughly corresponds to character-based decoding). Table 2 reports the number of encoding and decoding layers across the pre-trained Conformers.

Table 1 Summary of the three transformer-based automatic speech recognisers used as the basis of our experiment

Name	Language	Version	Parameters	Training set
Small [92]	English	1.6.0	∼ 14 M	ASRSET 2.0
Medium [93]	English	1.6.0	∼ 32 M	ASRSET 2.0
Large [94]	English	1.6.0	∼ 120 M	ASRSET 2.0

Table 2 Number of layers and neurons-per-network in the encoder and decoder modules of the three Transformer Automatic Speech Recognisers probed

Model	Encoder		Decoder			
	Layers	Neurons	Predictor		Joiner	
			Layers	Neurons	Layers	Neurons
Small	16	176	1	320	1	320
Medium	16	256	1	640	1	640
Large	17	512	1	640	1	640

In the present experiment, we probed the encoding layers' emissions of the three Nvidia Nemo Conformers (hereafter generally indicated as *Transformer ASR models*) to search for evidence of syllable-related information automatically forming in these layers.

3.2 Syllable boundary detection

To investigate whether the Transformer ASR models internally formed a rhythmic or syllabic-scale component in specific encoding layers, we trained new machine-learning models with the emission vectors of different encoding layers. Each model was trained to classify one emission vector from one encoding layer as corresponding to syllabic boundary *presence* or *absence*. We trained one detector for each encoding layer and Transformer ASR model size to verify (i) whether specific encoding layers contained sufficient information for syllable boundary detection and (ii) whether the correct detections could be associated with long and intense (prominent) syllables, which are integral to human speech recognition [11–16].

We used new training and test material to build the syllable boundary detectors (Sect. 3.3). The Transformer ASR models were preliminarily executed on the training set audio files to produce word-level transcriptions. Probing was conducted by acquiring tensors (later flattened into vectors) at the output of one encoding layer at a time. Specifically, all emission vectors $\{e_{l,m}\}$ (of length h), belonging to the probing of the l-th layer (between 1 and 16 or 17) of Transformer m (with m among Small, Medium, and Large) were saved as the training data for a new syllable boundary detection model.

The syllable boundary detection model was a Long Short-Term Memory (LSTM) model followed by a binary-classifying Feed-Forward artificial neural network (Fig. 3). LSTMs are suited models for classifying a time series of observation vectors [95] and making predictions out of

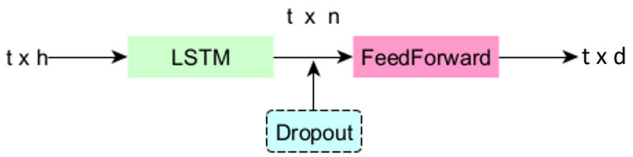

Fig. 3 Schema of our syllable boundary detector: t indicates the temporal index of the currently analysed frame; n is the LSTM hidden-layer length; h is the probed Transducer model's emission-vector length; d is the binary decision (0/1) produced by a classification feed-forward artificial neural network

historical data [96–99]. In the conventional architecture, an LSTM consists of one computational unit that iteratively processes all input time series vectors. This unit comprises three *gates* that process one vector at a time while combining this vector with information extracted from the previously processed vectors. All gates are realised as one-layer Feed-Forward neural networks with the same number of output neurons (hidden-layer length, n) and tanh or sigmoid activation functions. The gates' outputs are further processed by an *output* gate that produces an output vector with size n for the input vector processed at time t. The LSTM hidden-layer length is the crucial model parameter to optimise for gaining optimal classification performance. Our LSTM processed a sequence of $\{e_{l,m}\}$ emission vectors (each of length h) and produced a new sequence of vectors with size n. The two sequences were aligned over time. For each time step t, the Feed-Forward network produced a binary decision for syllable boundary presence (1) or absence (0) based on the LSTM hidden-layer output. In summary, we trained and tested different LSTM-based syllabic boundary detectors ($L_{n,m,l}$) for all possible n, m, and l combinations and studied the models' performance while searching for evidence of syllable-related properties in the models' decisions. To reduce overfitting risk, we also enabled a *dropout* neuron-selection strategy for the LSTM gates, which statistically excluded (with a 0.2 probability) one neuron and its weights at each training iteration [100]. Notably, our syllabic boundary detectors' temporal sensitivity (the minimum difference between consecutive time steps) was 40 ms because all Transformer ASR models produced emissions at this rate.

3.3 Experimental data

For testing the syllable boundary detectors' performance, we used a dataset annotated by Italian and Spanish experts, available for the members or customers of the CLARIN research infrastructure [101]. The Italian corpus contained 68 wave files from 11 speakers recorded at a 16-kHz sampling frequency for a total of 3.5 min of annotated audio. The Spanish corpus contained 45 recordings from 6 speakers for a total of 2.8 min of annotated audio.

Annotations were available in Textgrid format [102] and contained the following annotation levels:

- *Word*: the word-by-word orthophonic transcription of the speech signal;
- *Syllable-phonetic*: the phonetic pronunciation of the uttered syllables;
- *Syllable-phonologic*: the expected syllable transcription according to the word-level transcription and the phonologically predictable reduction processes.

These three annotation levels did not necessarily correspond to synchronised boundaries because they were produced independently from each other, and perceptive differences exist in the human recognition of the different levels [15]. In our experiment, we used the syllable-phonetic level to detect acoustic-related syllable boundaries. Acoustic-related syllables are indeed the only syllable types an ASR model can extract from the raw audio signal without using an externally provided linguistic model. Additionally, merging the Italian and Spanish corpora was plausible because these languages belong to the same linguistic family and have similar syllable boundary definitions. The merged corpus allowed us to produce statistically significant results.

We used the Italian-Spanish merged corpus to train and test our LSTM-based syllable boundary detectors. We split the corpus into train, validation, and test sets using 60, 20, and 20% percentages while ensuring that these sets did not share the same speakers. This choice aimed at guaranteeing that results were mostly speaker-independent. In the data preparation phase, we associated the Transformer ASR models' emissions with syllable boundary presence or absence and then used this association for model development and testing. Therefore, we prepared separate vector datasets for each probed emission layer of each Transformer ASR model. Based on these data, we conducted a two-step analysis: First, we detected the three most promising parametrisations of the LSTM-based syllable boundary detectors. These models were selected as having very different lengths of the LSTM hidden layer and achieving comparable high performance on the validation data. They allowed us to study performance variation across different resolutions of emissions' encoding and processing in the LSTMs. Second, we compared the syllable boundary detectors' performance across the Transformer ASR models and encoding layers. We also tested whether syllabic information detection was independent of emission encoding resolution in the LSTM.

Using Italian and Spanish corpora to train and test the syllable boundary detectors—although the Transformer ASR models were originally trained in English—was a reasoned experimental choice. Indeed, our target was to study rhythmic, syllable-related information rather than

syllable recognition. Rhythm is a language-independent feature, whereas syllable recognition is a language-dependent one [103]. Therefore, using languages other than English in the probing task allowed us to study the language-independent features of the syllabic and rhythmic components modelled by the Transformer ASR models.

4 Results

This section reports the performance measurements of our LSTM-based syllable boundary detectors across the probed encoding layers of the Transformer ASR models. In particular, Sect. 4.1 reports standard evaluation metrics for detecting the optimal LSTM parametrisation per Transformer ASR model, and Sect. 4.2 reports a syllable-oriented acoustic characterisation of the detected syllable boundaries.

4.1 Error rate and statistical significance

4.1.1 Considered metrics and measurements

We evaluated all possible $L_{n,m,l}$ models to identify the optimal LSTM hidden-layer length (n) and encoding layer depth (l) per Transformer ASR model size (m). The three best models achieving an overall high performance, with sufficiently different n, had the following hidden-layer lengths: 160, 320, and 740. We compared these models across all m and l combinations for a total of 172 L models trained and tested.

We used the SCTK[1] evaluation suite of the National Institute of Standards Technologies (NIST), a commonly used reference tool, to measure the $L_{n,m,l}$ models' performance. In particular, in compliance with other syllable boundary detectors [104–106], we measured the model Word Correct Rate (WCR) as the fraction of correctly classified words (i.e. $\text{WCR} = \frac{\text{Number of correctly classified words}}{\text{Total words}}$). The Word Error Rate metric, commonly used by other works, corresponds to $1 - \text{WCR}$. In our experiments, *word* corresponds to a syllable-boundary label indicating *presence* or *absence* in a 40-ms segment.

4.1.2 Evaluation

Figure 4a–c reports the WCR charts grouped by LSTM hidden-layer length. The x-axis indicates the probed Transformer ASR model's layer depth index, and the colours indicate the three Transformer ASR models analysed. The y-axis reports the WCR. For example, in Fig. 4a, $x = 0$ compares the syllable boundary detector with $n = 160$, trained on the emissions extracted from the first layer ($l = 1$) of the Small, Medium, and Large Transformer ASR models separately (i.e. for all m values).

The general trend emerging from the charts is that lower encoding layers contained higher discriminant information for syllable boundary detection, which decreased in deeper layers. Layers with depth indexes between 3 and 6 contained the most valuable information, with the 4th and 5th depth-index layers being the most informative. This observation was valid across all Transformer ASR models. The detectors' WCRs were more similar across the Transformer ASR models as far as the LSTM hidden-layer length increased. This observation indicates similar information encoding in 'long' LSTMs, which compensated for smaller Transformer ASR model sizes (Fig. 4c).

We also tested the statistical significance of the measured performance differences. After fixing m and n, we cross-compared the $(L_{n,m,l_i}, L_{n,m,l_j})$ WCRs for all i and j layers (with $i \neq j$). Significance tests were two-tailed tests with the null hypothesis that there was no significant WCR difference. For example, Table 3 reports all significance tests for a syllable boundary detector with a 320 hidden-layer length using the emissions of the Small Transformer ASR model (all other tables are reported in Appendix). Columns Sys 1 and Sys 2 indicate the l_i and l_j indices. The *Win* column indicates which detector achieved the highest performance. The *Relevance* column classifies the minimum significance level p-value as '*' ($p = 0.001$), '**' ($p = 0.01$), '***' ($p = 0.05$), or non-significant (empty). Therefore, the most significant discrepancies were those indicated with one '*'. The table demonstrates that the LSTM with a 320 hidden-layer length achieved the highest performance using the emissions of the 4th-index encoding layer of the Small Transformer ASR model. This performance was significantly higher than the one achieved using the other encoding layers. The comparisons across all n and m values confirmed that the 4th and 5th Transformer ASR model's encoding layer indexes always corresponded to the highest and most significant WCRs.

4.2 Optimal model identification and energy-pitch characterisation of the classifications

4.2.1 Considered metrics and measurements

We measured the overall performance of the $L_{n,m,l}$ models after fixing l to the most informative emission layer for syllable boundary detection per (n, m) pair. We used standard measurements (Accuracy, Precision, Recall, F1) based on the experts' corpus annotations. We also used

[1] https://github.com/usnistgov/SCTK.

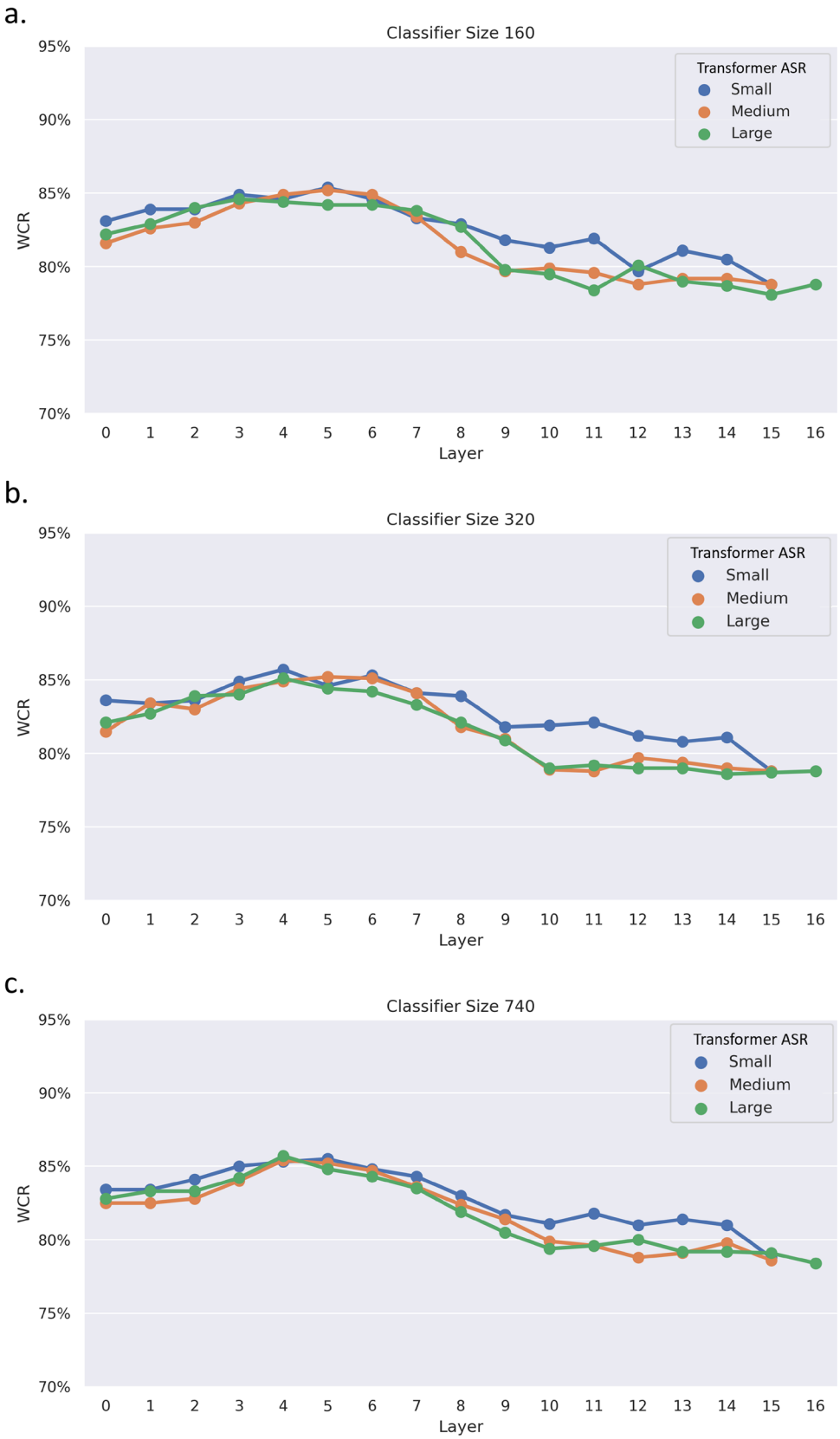

Fig. 4 Word correct rate of our LSTM-based syllable boundary detectors across the emission vectors extracted from three Transformer ASR models (small, medium, and large). The *x*-axis reports the depth of the Transformer ASR model layer from which vectors were extracted. The three charts correspond to different LSTM hidden-layer lengths, i.e. **a** 160, **b** 320, and **c** 740

Table 3 Summary of the pairwise statistical significance tests between LSTM-based syllable boundary detectors with a 320 hidden layer length, trained on the features vectors extracted from the Small Transducer ASR model

Sys 1	Sys 2	Win	Relevance	Sys 1	Sys 2	Win	Relevance	Sys 1	Sys 2	Win	Relevance
0	1			2	14	2	**	6	12	6	***
0	2			2	15	2	***	6	13	6	***
0	3			3	4			6	14	6	***
0	4	4	*	3	5			6	15	6	***
0	5			3	6			7	8		
0	6			3	7			7	9	7	***
0	7			3	8			7	10	7	**
0	8			3	9	3	***	7	11	7	***
0	9			3	10	3	**	7	12	7	***
0	10			3	11	3	***	7	13	7	***
0	11			3	12	3	***	7	14	7	***
0	12			3	13	3	***	7	15	7	***
0	13	0	*	3	14	3	***	8	9	8	***
0	14	0	*	3	15	3	***	8	10	8	**
0	15	0	***	4	5	4	*	8	11	8	**
1	2			4	6			8	12	8	***
1	3	3	*	4	7			8	13	8	***
1	4	4	**	4	8	4	*	8	14	8	***
1	5			4	9	4	***	8	15	8	***
1	6	6	*	4	10	4	***	9	10		
1	7			4	11	4	***	9	11		
1	8			4	12	4	***	9	12		
1	9			4	13	4	***	9	13		
1	10			4	14	4	***	9	14		
1	11			4	15	4	***	9	15	9	***
1	12			5	6			10	11		
1	13	1	*	5	7			10	12		
1	14	1	*	5	8			10	13		
1	15	1	***	5	9	5	**	10	14		
2	3			5	10	5	**	10	15	10	***
2	4	4	**	5	11	5	**	11	12		
2	5			5	12	5	**	11	13		
2	6			5	13	5	***	11	14		
2	7			5	14	5	***	11	15	11	***
2	8			5	15	5	***	12	13		
2	9	2	*	6	7			12	14		
2	10			6	8			12	15	12	***
2	11	2	*	6	9	6	***	13	14		
2	12	2	*	6	10	6	***	13	15	13	**
2	13	2	**	6	11	6	***	14	15	14	**

Cohen's kappa to measure the agreement between the manual and automatic annotations with respect to the chance agreement. In this comparison, true positives (TPs) were 40-ms segments where both the manual and automatic annotations indicated the presence of a syllabic boundary. Likewise, true negatives (TNs) were segments where both the annotations indicated the absence of a syllabic boundary. False negatives (FNs) were segments where only the manual annotation indicated a syllabic-boundary presence. Finally, false positives (FPs) were segments where only the automatic annotation indicated a syllabic-boundary presence. In summary, the following performance measurements were used:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{(\text{Precision} + \text{Recall})}$$

4.2.2 Evaluation

We used Accuracy and F1 as the principal measurements to identify the optimal model because Accuracy calculates the overall fraction of correctly detected boundaries, and F1 summarises Precision and Recall through their harmonic mean. Generally, high Accuracy was measured for all detectors, i.e. they could extract valuable syllable boundary-related information. The assessment indicated that the overall optimal model was an LSTM-based model with a 320 hidden layer length, operating on the output of the 4th layer index of the Small Transformer ASR model (Table 4). The kappa agreement was "good" according to Fleiss' classifications [107] for all detectors but was slightly better for the optimal model (0.54). The optimal model achieved a lower Recall than the other models because of a higher number of false positives. However, the model compensated for the Recall loss with a higher Precision, resulting in a higher F1 overall.

As an additional step, we characterised the optimal model's classification categories (TP–TN–FP–FN) over 40-ms segments by studying their average energy and pitch-level distributions. Energy is here intended as the squared sum of the signal-segment samples divided by the total number of segment samples (signal-segment *power*). Pitch, a rhythm-related feature, was estimated as the average pitch of 10-ms windows within the 40-ms classified segments. It was calculated through Boersma's sound-to-pitch algorithm [108] within a 60–250-Hz frequency band. The energy and pitch distributions across the classification categories allowed for characterising specific and shared properties of these categories (Table 5 and Fig. 5). Generally, the TPs corresponded to segments with higher energy than FPs (+ 31%) and TN (+ 14%) but had slightly lower energy than FNs (− 6%) (Table 5). Higher energy (+ 25%) was also observable for expert-annotated syllabic boundaries (corresponding to TP + FN) compared to non-annotated segments (TN + FP). Conversely, TPs corresponded to an averagely lower pitch than FPs (− 1%) but an averagely higher pitch than TNs (+ 18%) and FNs (+ 7%). However, the experts' annotations presented an average higher pitch (+ 5%) in the syllabic-boundary segments than in non-syllabic-boundary segments. Notably, TPs fell in energy islands (signal segments characterised by an increasing onset, a nucleus, and a decreasing coda) of syllabic scale (100–200 ms) and with averagely double the duration of the TNs' energy islands (40–100 ms).

A range of energy values over 34.4E5 mainly corresponded to TPs (Fig. 5a), which would enforce the classification confidence of these segments as syllabic boundaries, should energy be used as a weighting classification factor. FNs presented moderately high energy (11.48E5 median) and pitch (124.29 Hz median and 157.53 Hz at the 75th percentile) (Fig. 5b). High pitch in FNs was a distinctive characteristic compared to TNs (119.66 Hz median value and 147.70 Hz at the 75th percentile) that would allow for automatically revising the classification of non-syllabic boundaries. As for FPs, the corresponding segments presented a median energy comparable with the TNs' energy (9.99E5 vs 9.64E5) but had lower median energy than the FNs (11.48E5). Therefore, energy was not a discriminant property of FPs. However, the FPs presented a generally higher pitch than TNs (129.64 Hz vs 119.66 Hz), which could help detect and correct some FPs.

5 Discussion and conclusions

This paper has described a probing experiment for end-to-end Transformer ASR models based on automatic syllable boundary detection. Our goal was to verify if such architectures internally modelled a rhythmic component similar to what humans appear to do while processing speech. Syllable boundary detection was based on an LSTM processing the feature vectors extracted from a Transformer ASR model's encoding layer. The most informative vectors were produced by the smallest-size Transformer ASR model and were optimally recognised through an LSTM with a 320 hidden-state length (medium size). Our syllable boundary detector also reached a higher accuracy ($\sim$ 87%) than alternative systems for Italian [104, 109, 110].

One significant result of the present study is that our syllable boundary detectors' performance depended on a rhythmic component modelled by the inner layers of the analysed Transformers, correlated with psycho-acoustic syllables. In fact, our evaluation highlighted that an acoustic component with high energy and long duration was primarily contained in the shallower Transformer's encoding layers ($\sim$ 4), fading out in deeper layers ($\sim$ 16), and was valuable for automatic syllable boundary detection. This result suggests that the Transformer ASR models captured syllable separation (and rhythm, consequently) in the earliest stage of the encoding process, in agreement

Table 4 Summary of the performance of our syllable boundary detectors reported per LSTM hidden-layer length. Each row reports the corresponding Transformer ASR model used and the optimal encoding layer used for feature extraction. Red numbers indicate the highest values for each measurement. The overall optimal model is highlighted in green

LSTM hidden layer length	Transformer's size	Transformer's optimal layer depth	Accuracy	Precision	Recall	F1	TP	TN	FP	FN	Cohen's Kappa
160	Large	4	85.84	66.79	57.09	61.56	169	1111	84	127	0.53
320	Large	4	86.11	68.46	55.74	61.44	165	1119	76	131	0.53
740	Large	4	86.65	71.55	54.39	61.8	161	1131	64	135	0.53
160	Medium	5	86.25	70.4	53.04	60.49	157	1129	66	139	0.52
320	Medium	4	86.31	70.35	53.71	60.91	159	1128	67	137	0.52
740	Medium	5	86.58	71.42	54.05	61.53	160	1131	64	136	0.53
160	Small	5	86.72	72.07	54.05	61.77	160	1133	62	136	0.53
320	Small	4	86.85	72.12	55.06	62.44	163	1132	63	133	0.54
740	Small	5	86.31	70.35	53.71	60.91	159	1128	67	137	0.52

Table 5 Proportions and relative variations of average energy and pitch in 40-ms length audio segments

Energy and pitch proportions and relative variations		
	Ratio	Relative variation
Average TP Energy/Average FP Energy	1.44	31%
Average TP Energy/Average TN Energy	1.17	14%
Average TP Energy/Average FN Energy	0.95	− 6%
Average TP + FN Energy/Average TN + FP Energy	1.33	25%
Average TP Pitch/Average FP Pitch	0.99	− 1%
Average TP Pitch/Average TN Pitch	1.22	18%
Average TP Pitch/Average FN Pitch	1.07	7%
Average TP + FN Pitch/Average TN + FP Pitch	1.05	5%

The table compares the values of the true positive classifications (TP) with those of the false positive (FP), true negative (TN), and false negative (FN) classifications

with studies that have explored automatic and human speech-processing similarities from a medical perspective [111]. It also aligns with other studies [76] that detected phone-level and word-level positive reactions in layers likely compatible with those we detected as reacting to syllables. A detailed analysis of the optimal syllable boundary detector output indicated that the true-positive classifications were associated with highly energetic boundaries within syllabic-scale energy islands (100–200 ms) having double the duration of the true negatives' duration. This observation indicates a correlation between the detected syllable boundaries and syllabic prominence [15]. The wrongly classified boundaries (false positives) had similar medium–low energy profiles to true negatives but an averagely higher pitch. Therefore, a high pitch and medium–low energy segment should lower the syllable boundary detection confidence, whereas a high energy segment should increase the detection confidence [104, 110].

Among the missed boundaries (false negatives), a subset was characterised by higher energy and pitch than true negatives. These cases might correspond to the boundaries of stressed syllables at the end of words (present in Italian and Spanish). Therefore, they could be due to the discrepancy between the Transformer ASR model and the syllable detector training languages. One point of discussion is indeed the consequence of training the Transformer in English and the syllable boundary detector in Italian and Spanish. The representation formed in the Transformer's shallower layers was related to English syllables, i.e. to the specific energy, length, and pitch profiles of a stress-timed language. Conversely, Italian and Spanish are syllable-timed languages. This discrepancy mainly increased the number of false negatives, although not enough to compromise the overall performance. The underlying reason

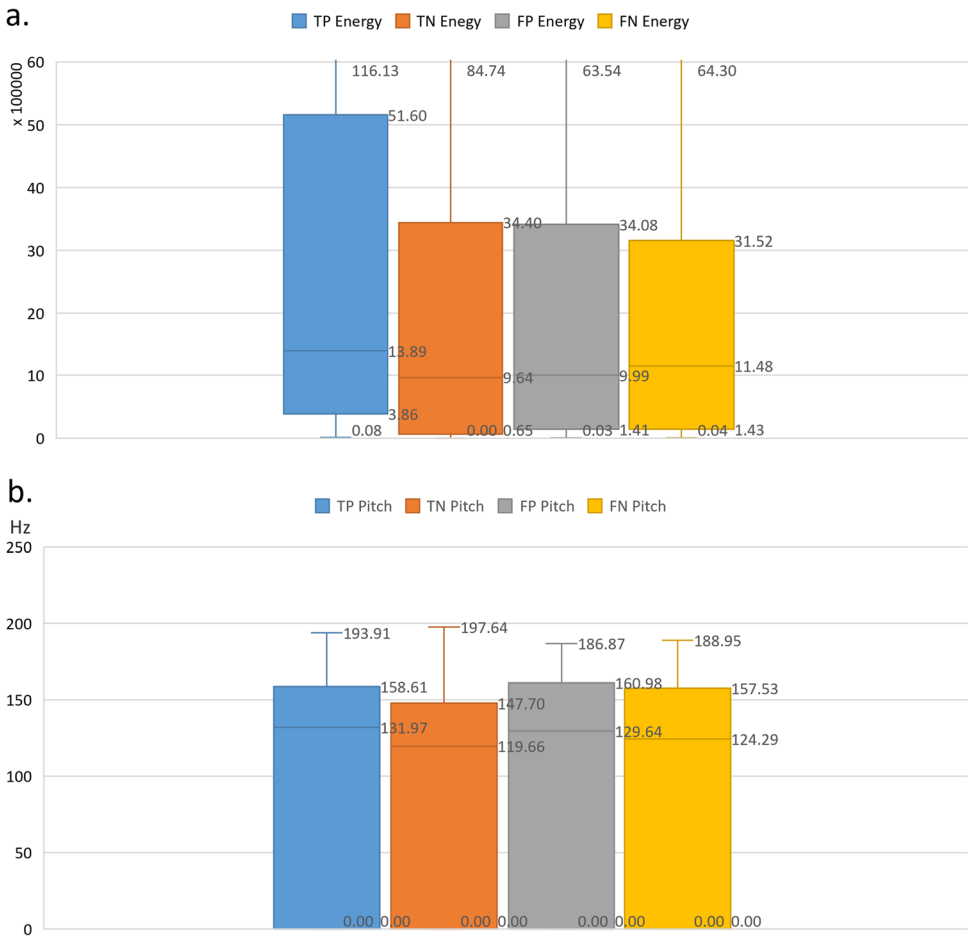

Fig. 5 Box plots displaying the distributions of **a** energy and **b** pitch values across true positive (TP), true negative (TN), false positive (FP), and false negative (FN) classifications

was likely that the emissions contained an important rhythmic component that was language-independent.

In this work, we have focused on syllable units rather than phonemes or words since the size of the considered analysis windows does not allow for capturing phonemic-related characteristics [76]. On the other hand, the encoding layer does not contain word-level information. In the future, we will consider finer- and coarse-grained units for analysis in more extended models. Moreover, we will explore how performance might change when all training sets belong to the same language. We will also study whether, in these conditions, true positives mostly correspond to long and intense syllables (i.e. to syllabic acoustic prominence). Having a way to detect prominent syllables would be crucial to improve syllabic ASR models' performance while drastically reducing the training set dimension [5] and would help refine the perceptive and acoustic definition of syllable [15].

Our results create an interesting parallel between human speech recognition, relying on psycho-acoustic syllable-related units, and DL-ASR internal processing. They provide insight and location about human-explainable processes inside E2E-ASR systems related to the formation of syllabic-scale unit representations. Other scientific studies have also conjectured that the internal knowledge representation formed in deep learning models can produce new definitions of speech units and emerging dynamics similar to the internal human brain's speech representations [112, 113]. However, it is difficult to understand the influence of the spontaneously formed speech units on the E2E-ASR performance due to the large number of parameters, training material, and the diverse training methodologies used [22, 114]. One common research question in this context is if we can learn from the psycho-acoustic-like dynamics in ASRs to enhance other ASRs' performance and efficiency. This question has been investigated in under-represented languages, limited-vocabulary ASRs, and robust spontaneous speech recognition in noisy environments [13, 22, 28, 115–117]. The question gains more interest if we highlight the gap between conventional ASRs (which use an explicit speech-unit and language modelling) and E2E-ASRs as the number of parameters used by the two system types. For example, the Wav2Vec2.0 E2E-ASR requires 10^8 parameters to

overcome the performance of a conventional ASR (with 10^2 parameters) on a large-vocabulary recognition task [114]. Performance is lower than conventional ASRs' performance when using 10^7 parameters. The combination of low complexity and high performance in conventional ASRs is due to their explicit modelling approach. However, new solutions might exist in the 10^6 gap of used parameters, which could rely on deep learning architectures using information from conventional ASR modelling [5]. Multi-channel E2E-ASRs are exploring this possibility by injecting supra-segmental and non-verbal characteristics in encoding layers to enhance noisy-speech recognition [118, 119]. They usually represent these characteristics as additional (latent) variables or pre-trained sub-models [120, 121]. Other approaches use the vectors extracted from the hidden layers of large E2E-ASRs (distilled features) to train smaller ASRs and achieve higher performance on specific tasks [122]. Conversely, other systems use distilled features instead of standard acoustic features to improve the performance of conventional ASRs [33, 123, 124].

Our experiment identified a particular type of distilled features related to rhythm and syllables that can be used in other ASRs. These features are suitable for Few-shot Learning, i.e. to make an ASR model generalise over new data categories using limited training data [125]. Distilled features similar to the ones we detected have indeed shown potential to reduce the hypothesis space, avoid overfitting, ensure heterogeneity in the prediction space, and consequently improve ASR effectiveness over the small datasets available for low-resourced languages and applications [126]. For example, they have been used as prototype vectors for internal encoding classes (e.g. speech units) to enhance class centroid representations and achieve better generalisation [127–129]. Moreover, they have been proposed to focus a Few-shot Learning model on *islands* of prominent-speech segments having high-quality pronunciation and thus being more clearly recognisable [5, 130].

Generally, ASR performance improvement deriving from integrating syllabic-scale features has been long reported and inspired the present work. Acoustic features enriched with syllable-boundary information or syllabic-scale features can sensibly improve continuous and spontaneous automatic speech recognition, especially in high noise and reverb scenarios [131–134]. Moreover, syllabic-scale features derived from deep learning models are critical for diagnostic systems based on prosodic information, such as those for pathological speech detection in syllable-timed languages [135, 136]. Recent studies have also highlighted the centrality of these features in contexts where prosody is the primary information source, such as infant cry detection and classification [137, 138]. The highly prosodic nature of infant cry indeed makes syllable-scale acoustic features central for these tasks, especially when extracted through deep learning models [99], and allows interpreting a newborn's psychological and clinical status [139–141]. Finally, another field of application of syllabic-scale features is the improvement of ASR robustness to adversarial attacks (e.g. hidden voice commands), which requires introducing new paradigms for attack evaluation [142–144]. Some studies have indeed highlighted that syllabification (which can be based on syllabic-scale features, like in our case) is critical to discovering potential attacks and consequently improving ASR robustness [145, 146].

In summary, all mentioned application cases would likely benefit from syllable-related distilled features, e.g. those extracted from the 4th-index encoding layer of the present paper's 320 hidden-state Transformer ASR. In future experiments, we will verify this statement in all mentioned contexts.

Appendix

See Tables 6, 7, 8, 9, 10, 11, 12, 13, and 14.

Table 6 Summary of the pairwise statistical significance tests between LSTM-based syllable classifiers with a 160 hidden layer length, trained on the features vectors extracted from the Small Transducer ASR

Sys 1	Sys 2	Win	Relevance	Sys 1	Sys 2	Win	Relevance	Sys 1	Sys 2	Win	Relevance
0	1			2	14	2	***	6	12	6	***
0	2			2	15	2	***	6	13	6	***
0	3	3	*	3	4			6	14	6	***
0	4			3	5			6	15	6	***
0	5	5	**	3	6			7	8		
0	6			3	7			7	9	7	*
0	7			3	8	3	*	7	10	7	**
0	8			3	9	3	**	7	11	7	*
0	9			3	10	3	***	7	12	7	***
0	10			3	11	3	**	7	13	7	**
0	11			3	12	3	***	7	14	7	***
0	12	0	**	3	13	3	***	7	15	7	***
0	13	0	*	3	14	3	***	8	9		
0	14	0	*	3	15	3	***	8	10	8	*
0	15	0	***	4	5			8	11		
1	2			4	6			8	12	8	***
1	3			4	7			8	13	8	*
1	4			4	8			8	14	8	**
1	5	5	**	4	9	4	**	8	15	8	***
1	6			4	10	4	***	9	10		
1	7			4	11	4	***	9	11		
1	8			4	12	4	***	9	12	9	**
1	9			4	13	4	***	9	13		
1	10	1	**	4	14	4	***	9	14		
1	11	1	*	4	15	4	***	9	15	9	***
1	12	1	***	5	6			10	11		
1	13	1	**	5	7	5	**	10	12		
1	14	1	**	5	8	5	***	10	13		
1	15	1	***	5	9	5	***	10	14		
2	3			5	10	5	***	10	15	10	**
2	4			5	11	5	***	11	12	11	**
2	5	5	*	5	12	5	***	11	13		
2	6			5	13	5	***	11	14		
2	7			5	14	5	***	11	15	11	***
2	8			5	15	5	***	12	13		
2	9	2	*	6	7			12	14		
2	10	2	**	6	8	6	*	12	15		
2	11	2	*	6	9	6	**	13	14		
2	12	2	***	6	10	6	***	13	15	13	**
2	13	2	**	6	11	6	***	14	15	14	**

Table 7 Summary of the pairwise statistical significance tests between LSTM-based syllable classifiers with a 320 hidden layer length, trained on the features vectors extracted from the Small Transducer ASR

Sys 1	Sys 2	Win	Relevance	Sys 1	Sys 2	Win	Relevance	Sys 1	Sys 2	Win	Relevance
0	1			2	14	2	**	6	12	6	***
0	2			2	15	2	***	6	13	6	***
0	3			3	4			6	14	6	***
0	4	4	*	3	5			6	15	6	***
0	5			3	6			7	8		
0	6			3	7			7	9	7	***
0	7			3	8			7	10	7	**
0	8			3	9	3	***	7	11	7	***
0	9			3	10	3	**	7	12	7	***
0	10			3	11	3	***	7	13	7	***
0	11			3	12	3	***	7	14	7	***
0	12			3	13	3	***	7	15	7	***
0	13	0	*	3	14	3	***	8	9	8	***
0	14	0	*	3	15	3	***	8	10	8	**
0	15	0	***	4	5	4	*	8	11	8	**
1	2			4	6			8	12	8	***
1	3	3	*	4	7			8	13	8	***
1	4	4	**	4	8	4	*	8	14	8	***
1	5			4	9	4	***	8	15	8	***
1	6	6	*	4	10	4	***	9	10		
1	7			4	11	4	***	9	11		
1	8			4	12	4	***	9	12		
1	9			4	13	4	***	9	13		
1	10			4	14	4	***	9	14		
1	11			4	15	4	***	9	15	9	***
1	12			5	6			10	11		
1	13	1	*	5	7			10	12		
1	14	1	*	5	8			10	13		
1	15	1	***	5	9	5	**	10	14		
2	3			5	10	5	**	10	15	10	***
2	4	4	**	5	11	5	**	11	12		
2	5			5	12	5	**	11	13		
2	6			5	13	5	***	11	14		
2	7			5	14	5	***	11	15	11	***
2	8			5	15	5	***	12	13		
2	9	2	*	6	7			12	14		
2	10			6	8			12	15	12	***
2	11	2	*	6	9	6	***	13	14		
2	12	2	*	6	10	6	***	13	15	13	**
2	13	2	**	6	11	6	***	14	15	14	**

Table 8 Summary of the pairwise statistical significance tests between LSTM-based syllable classifiers with a 740 hidden layer length, trained on the features vectors extracted from the Small Transducer ASR

Sys 1	Sys 2	Win	Relevance	Sys 1	Sys 2	Win	Relevance	Sys 1	Sys 2	Win	Relevance
0	1			2	14	2	***	6	12	6	***
0	2			2	15	2	***	6	13	6	***
0	3	3	*	3	4			6	14	6	***
0	4	4	*	3	5			6	15	6	***
0	5	5	*	3	6			7	8	7	*
0	6			3	7			7	9	7	***
0	7			3	8			7	10	7	***
0	8			3	9	3	**	7	11	7	***
0	9			3	10	3	**	7	12	7	***
0	10			3	11	3	**	7	13	7	***
0	11			3	12	3	***	7	14	7	***
0	12	0	*	3	13	3	***	7	15	7	***
0	13			3	14	3	***	8	9	8	*
0	14	0	*	3	15	3	***	8	10	8	*
0	15	0	***	4	5			8	11		
1	2	2	*	4	6			8	12	8	*
1	3	3	**	4	7			8	13	8	*
1	4	4	**	4	8	4	*	8	14	8	*
1	5	5	**	4	9	4	***	8	15	8	***
1	6	6	*	4	10	4	***	9	10		
1	7	7	*	4	11	4	***	9	11		
1	8			4	12	4	***	9	12		
1	9			4	13	4	***	9	13		
1	10			4	14	4	***	9	14		
1	11			4	15	4	***	9	15	9	***
1	12			5	6			10	11		
1	13			5	7			10	12		
1	14	1	*	5	8	5	**	10	13		
1	15	1	***	5	9	5	***	10	14		
2	3			5	10	5	***	10	15	10	**
2	4			5	11	5	***	11	12		
2	5			5	12	5	***	11	13		
2	6			5	13	5	***	11	14		
2	7			5	14	5	***	11	15	11	***
2	8			5	15	5	***	12	13		
2	9	2	**	6	7			12	14		
2	10	2	**	6	8	6	*	12	15	12	**
2	11	2	**	6	9	6	***	13	14		
2	12	2	***	6	10	6	***	13	15	13	***
2	13	2	***	6	11	6	**	14	15	14	**

Table 9 Summary of the pairwise statistical significance tests between LSTM-based syllable classifiers with a 160 hidden layer length, trained on the features vectors extracted from the Medium Transducer ASR

Sys 1	Sys 2	Win	Relevance	Sys 1	Sys 2	Win	Relevance	Sys 1	Sys 2	Win	Relevance
0	1			2	14	2	***	6	12	6	***
0	2			2	15	2	***	6	13	6	***
0	3	3	***	3	4			6	14	6	***
0	4	4	***	3	5			6	15	6	***
0	5	5	***	3	6			7	8	7	**
0	6	6	***	3	7			7	9	7	***
0	7	7	*	3	8	3	**	7	10	7	***
0	8			3	9	3	***	7	11	7	***
0	9			3	10	3	***	7	12	7	***
0	10			3	11	3	***	7	13	7	***
0	11			3	12	3	***	7	14	7	***
0	12	0	**	3	13	3	***	7	15	7	***
0	13			3	14	3	***	8	9		
0	14	0	*	3	15	3	***	8	10		
0	15	0	**	4	5			8	11	8	*
1	2			4	6			8	12	8	**
1	3	3	**	4	7			8	13	8	*
1	4	4	**	4	8	4	***	8	14	8	**
1	5	5	**	4	9	4	***	8	15	8	**
1	6	6	**	4	10	4	***	9	10		
1	7			4	11	4	***	9	11		
1	8			4	12	4	***	9	12		
1	9	1	*	4	13	4	***	9	13		
1	10	1	*	4	14	4	***	9	14		
1	11	1	**	4	15	4	***	9	15		
1	12	1	***	5	6			10	11		
1	13	1	**	5	7			10	12		
1	14	1	***	5	8	5	***	10	13		
1	15	1	***	5	9	5	***	10	14		
2	3	3	**	5	10	5	***	10	15		
2	4	4	**	5	11	5	***	11	12		
2	5	5	**	5	12	5	***	11	13		
2	6	6	**	5	13	5	***	11	14		
2	7			5	14	5	***	11	15		
2	8			5	15	5	***	12	13		
2	9	2	**	6	7			12	14		
2	10	2	**	6	8	6	***	12	15		
2	11	2	***	6	9	6	***	13	14		
2	12	2	***	6	10	6	***	13	15		
2	13	2	***	6	11	6	***	14	15		

Table 10 Summary of the pairwise statistical significance tests between LSTM-based syllable classifiers with a 320 hidden layer length, trained on the features vectors extracted from the Medium Transducer ASR

Sys 1	Sys 2	Win	Relevance	Sys 1	Sys 2	Win	Relevance	Sys 1	Sys 2	Win	Relevance
0	1	1	**	2	14	2	***	6	12	6	***
0	2	2	*	2	15	2	***	6	13	6	***
0	3	3	***	3	4			6	14	6	***
0	4	4	***	3	5			6	15	6	***
0	5	5	***	3	6			7	8	7	***
0	6	6	***	3	7			7	9	7	***
0	7	7	**	3	8	3	**	7	10	7	***
0	8			3	9	3	***	7	11	7	***
0	9			3	10	3	***	7	12	7	***
0	10	0	*	3	11	3	***	7	13	7	***
0	11	0	*	3	12	3	***	7	14	7	***
0	12			3	13	3	***	7	15	7	***
0	13	0	*	3	14	3	***	8	9		
0	14	0	*	3	15	3	***	8	10	8	***
0	15	0	**	4	5			8	11	8	***
1	2			4	6			8	12	8	*
1	3			4	7			8	13	8	**
1	4	4	**	4	8	4	***	8	14	8	**
1	5	5	*	4	9	4	***	8	15	8	***
1	6	6	*	4	10	4	***	9	10	9	***
1	7			4	11	4	***	9	11	9	**
1	8			4	12	4	***	9	12	9	*
1	9	1	*	4	13	4	***	9	13	9	*
1	10	1	***	4	14	4	***	9	14	9	**
1	11	1	***	4	15	4	***	9	15	9	***
1	12	1	***	5	6			10	11		
1	13	1	***	5	7			10	12		
1	14	1	***	5	8	5	***	10	13		
1	15	1	***	5	9	5	***	10	14		
2	3	3	*	5	10	5	***	10	15		
2	4	4	**	5	11	5	***	11	12		
2	5	5	**	5	12	5	***	11	13		
2	6	6	**	5	13	5	***	11	14		
2	7			5	14	5	***	11	15		
2	8			5	15	5	***	12	13		
2	9	2	*	6	7			12	14		
2	10	2	***	6	8	6	***	12	15		
2	11	2	***	6	9	6	***	13	14		
2	12	2	**	6	10	6	***	13	15		
2	13	2	***	6	11	6	***	14	15		

Table 11 Summary of the pairwise statistical significance tests between LSTM-based syllable classifiers with a 740 hidden layer length, trained on the features vectors extracted from the Medium Transducer ASR

Sys 1	Sys 2	Win	Relevance	Sys 1	Sys 2	Win	Relevance	Sys 1	Sys 2	Win	Relevance
0	1			2	14	2	***	6	12	6	***
0	2			2	15	2	***	6	13	6	***
0	3	3	**	3	4	4	*	6	14	6	***
0	4	4	***	3	5			6	15	6	***
0	5	5	***	3	6			7	8	7	*
0	6	6	***	3	7			7	9	7	**
0	7			3	8			7	10	7	***
0	8			3	9	3	**	7	11	7	***
0	9			3	10	3	***	7	12	7	***
0	10	0	*	3	11	3	***	7	13	7	***
0	11	0	*	3	12	3	***	7	14	7	***
0	12	0	**	3	13	3	***	7	15	7	***
0	13	0	**	3	14	3	***	8	9		
0	14	0	*	3	15	3	***	8	10	8	***
0	15	0	***	4	5			8	11	8	***
1	2			4	6			8	12	8	***
1	3			4	7			8	13	8	***
1	4	4	***	4	8	4	**	8	14	8	***
1	5	5	**	4	9	4	***	8	15	8	***
1	6	6	**	4	10	4	***	9	10	9	*
1	7			4	11	4	***	9	11	9	**
1	8			4	12	4	***	9	12	9	***
1	9			4	13	4	***	9	13	9	**
1	10	1	**	4	14	4	***	9	14	9	**
1	11	1	**	4	15	4	***	9	15	9	***
1	12	1	***	5	6			10	11		
1	13	1	***	5	7	5	*	10	12	10	*
1	14	1	**	5	8	5	***	10	13		
1	15	1	***	5	9	5	***	10	14		
2	3			5	10	5	***	10	15	10	*
2	4	4	***	5	11	5	***	11	12		
2	5	5	***	5	12	5	***	11	13		
2	6	6	**	5	13	5	***	11	14		
2	7			5	14	5	***	11	15		
2	8			5	15	5	***	12	13		
2	9			6	7			12	14		
2	10	2	**	6	8	6	**	12	15		
2	11	2	***	6	9	6	***	13	14		
2	12	2	***	6	10	6	***	13	15		
2	13	2	***	6	11	6	***	14	15		

Table 12 Summary of the pairwise statistical significance tests between LSTM-based syllable classifiers with a 160 hidden layer length, trained on the features vectors extracted from the Large Transducer ASR

Sys 1	Sys 2	Win	Relevance	Sys 1	Sys 2	Win	Relevance	Sys 1	Sys 2	Win	Relevance	Sys 1	Sys 2	Win	Relevance
0	1			2	12	2	***	5	16	5	***	10	16		
0	2	2	**	2	13	2	***	6	7			11	12	12	*
0	3	3	***	2	14	2	***	6	8			11	13		
0	4	4	**	2	15	2	***	6	9	6	***	11	14		
0	5	5	**	2	16	2	***	6	10	6	***	11	15		
0	6	6	**	3	4			6	11	6	***	11	16		
0	7	7	*	3	5			6	12	6	***	12	13		
0	8			3	6			6	13	6	***	12	14		
0	9			3	7			6	14	6	***	12	15	12	**
0	10	0	*	3	8			6	15	6	***	12	16	12	*
0	11	0	***	3	9	3	***	6	16	6	***	13	14		
0	12	0	*	3	10	3	***	7	8			13	15		
0	13	0	**	3	11	3	***	7	9	7	***	13	16		
0	14	0	**	3	12	3	***	7	10	7	***	14	15		
0	15	0	***	3	13	3	***	7	11	7	***	14	16		
0	16	0	***	3	14	3	***	7	12	7	***	15	16		
1	2			3	15	3	***	7	13	7	***				
1	3	3	*	3	16	3	***	7	14	7	***				
1	4	4	*	4	5			7	15	7	***				
1	5			4	6			7	16	7	***				
1	6			4	7			8	9	8	***				
1	7			4	8			8	10	8	***				
1	8			4	9	4	***	8	11	8	***				
1	9	1	**	4	10	4	***	8	12	8	***				
1	10	1	***	4	11	4	***	8	13	8	***				
1	11	1	***	4	12	4	***	8	14	8	***				
1	12	1	***	4	13	4	***	8	15	8	***				
1	13	1	***	4	14	4	***	8	16	8	***				
1	14	1	***	4	15	4	***	9	10						
1	15	1	***	4	16	4	***	9	11	9	*				
1	16	1	***	5	6			9	12						
2	3			5	7			9	13						
2	4			5	8	5	*	9	14						
2	5			5	9	5	***	9	15	9	**				
2	6			5	10	5	***	9	16	9	*				
2	7			5	11	5	***	10	11						
2	8			5	12	5	***	10	12						
2	9	2	***	5	13	5	***	10	13						
2	10	2	***	5	14	5	***	10	14						
2	11	2	***	5	15	5	***	10	15	10	*				

Table 13 Summary of the pairwise statistical significance tests between LSTM-based syllable classifiers with a 320 hidden layer length, trained on the features vectors extracted from the Large Transducer ASR

Sys 1	Sys 2	Win	Relevance	Sys 1	Sys 2	Win	Relevance	Sys 1	Sys 2	Win	Relevance	Sys 1	Sys 2	Win	Relevance
0	1			2	12	2	***	5	16	5	***	10	16		
0	2	2	**	2	13	2	***	6	7			11	12		
0	3	3	**	2	14	2	***	6	8	6	*	11	13		
0	4	4	***	2	15	2	***	6	9	6	***	11	14		
0	5	5	**	2	16	2	***	6	10	6	***	11	15		
0	6	6	**	3	4			6	11	6	***	11	16		
0	7			3	5			6	12	6	***	12	13		
0	8			3	6			6	13	6	***	12	14		
0	9			3	7			6	14	6	***	12	15		
0	10	0	*	3	8	3	*	6	15	6	***	12	16		
0	11	0	*	3	9	3	***	6	16	6	***	13	14		
0	12	0	**	3	10	3	***	7	8			13	15		
0	13	0	**	3	11	3	***	7	9	7	***	13	16		
0	14	0	**	3	12	3	***	7	10	7	***	14	15		
0	15	0	**	3	13	3	***	7	11	7	***	14	16		
0	16	0	**	3	14	3	***	7	12	7	***	15	16		
1	2			3	15	3	***	7	13	7	***				
1	3			3	16	3	***	7	14	7	***				
1	4	4	**	4	5			7	15	7	***				
1	5	5	*	4	6			7	16	7	***				
1	6			4	7			8	9	8	**				
1	7			4	8	4	**	8	10	8	***				
1	8			4	9	4	***	8	11	8	***				
1	9	1	*	4	10	4	***	8	12	8	***				
1	10	1	***	4	11	4	***	8	13	8	***				
1	11	1	***	4	12	4	***	8	14	8	***				
1	12	1	***	4	13	4	***	8	15	8	***				
1	13	1	***	4	14	4	***	8	16	8	***				
1	14	1	***	4	15	4	***	9	10						
1	15	1	***	4	16	4	***	9	11	9	*				
1	16	1	***	5	6			9	12	9	**				
2	3			5	7			9	13	9	**				
2	4			5	8	5	**	9	14	9	**				
2	5			5	9	5	***	9	15	9	***				
2	6			5	10	5	***	9	16	9	**				
2	7			5	11	5	***	10	11						
2	8			5	12	5	***	10	12						
2	9	2	***	5	13	5	***	10	13						
2	10	2	***	5	14	5	***	10	14						
2	11	2	***	5	15	5	***	10	15						

Table 14 Summary of the pairwise statistical significance tests between LSTM-based syllable classifiers with a 740 hidden layer length, trained on the features vectors extracted from the Large Transducer ASR

Sys 1	Sys 2	Win	Relevance	Sys 1	Sys 2	Win	Relevance	Sys 1	Sys 2	Win	Relevance	Sys 1	Sys 2	Win	Relevance
0	1			2	12	2	***	5	16	5	***	10	16	10	*
0	2			2	13	2	***	6	7			11	12		
0	3	3	*	2	14	2	***	6	8	6	**	11	13		
0	4	4	***	2	15	2	***	6	9	6	***	11	14		
0	5	5	*	2	16	2	***	6	10	6	***	11	15		
0	6	6	*	3	4			6	11	6	***	11	16	11	*
0	7			3	5			6	12	6	***	12	13		
0	8			3	6			6	13	6	***	12	14		
0	9	0	*	3	7			6	14	6	***	12	15		
0	10	0	**	3	8	3	*	6	15	6	***	12	16	12	**
0	11	0	***	3	9	3	***	6	16	6	***	13	14		
0	12	0	**	3	10	3	***	7	8	7	*	13	15		
0	13	0	***	3	11	3	***	7	9	7	***	13	16		
0	14	0	***	3	12	3	***	7	10	7	***	14	15		
0	15	0	***	3	13	3	***	7	11	7	***	14	16	14	*
0	16	0	***	3	14	3	***	7	12	7	***	15	16	15	*
1	2			3	15	3	***	7	13	7	***				
1	3			3	16	3	***	7	14	7	***				
1	4	4	***	4	5			7	15	7	***				
1	5			4	6			7	16	7	***				
1	6			4	7	4	*	8	9	8	*				
1	7			4	8	4	***	8	10	8	***				
1	8			4	9	4	***	8	11	8	***				
1	9	1	*	4	10	4	***	8	12	8	***				
1	10	1	***	4	11	4	***	8	13	8	***				
1	11	1	***	4	12	4	***	8	14	8	***				
1	12	1	***	4	13	4	***	8	15	8	***				
1	13	1	***	4	14	4	***	8	16	8	***				
1	14	1	***	4	15	4	***	9	10						
1	15	1	***	4	16	4	***	9	11	9	*				
1	16	1	***	5	6			9	12						
2	3			5	7			9	13	9	**				
2	4	4	*	5	8	5	**	9	14	9	*				
2	5			5	9	5	***	9	15	9	*				
2	6			5	10	5	***	9	16	9	***				
2	7			5	11	5	***	10	11						
2	8			5	12	5	***	10	12						
2	9	2	**	5	13	5	***	10	13						
2	10	2	***	5	14	5	***	10	14						
2	11	2	***	5	15	5	***	10	15						

Funding Open access funding provided by ISTI - PISA within the CRUI-CARE Agreement. The paper is the result of a self-funded research initiative

Data availability Source code for syllabic unit boundary detection and the analysis results are publicly available and downloadable from the D4Science e-Infrastructure [147–149] at the following public link https://data.d4science.net/3Jk5.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethical approval Not applicable.

References

1. Lu H, Li Y, Chen M, Kim H, Serikawa S (2018) Brain intelligence: go beyond artificial intelligence. Mob Netw Appl 23(2):368–375
2. Kamath U, Liu J, Whitaker J (2019) Deep learning for NLP and speech recognition, vol 84. Springer, Heidelberg
3. Wang S, Li G (2019) Overview of end-to-end speech recognition. J Phys Conf Ser 1187:052068
4. Aggarwal S, Sharma S (2021) Voice based deep learning enabled user interface design for smart home application system. In: 2021 2nd International conference on communication, computing and industry 4.0 (C2I4). IEEE, pp 1–6
5. Coro G, Massoli FV, Origlia A, Cutugno F (2021) Psycho-acoustics inspired automatic speech recognition. Comput Electr Eng 93:107238
6. Grabe E, Low EL (2002) Durational variability in speech and the rhythm class hypothesis. Pap Lab Phonol 7(1982):515–546
7. Nokes J, Hay J (2012) Acoustic correlates of rhythm in New Zealand English: a diachronic study. Lang Var Chang 24(1):1–31
8. D'Alessandro C, Mertens P (1995) Automatic pitch contour stylization using a model of tonal perception. Comput Speech Lang 9(3):257–288
9. Roach P (2000) English phonetics and phonology. A practical course. Cambridge University Press, Cambridge
10. MacNeilage PF, Davis BL (2000) On the origin of internal structure of word forms. Science 288(5465):527–531
11. Massaro DW, Stork DG (1998) Speech recognition and sensory integration: a 240-year-old theorem helps explain how people and machines can integrate auditory and visual information to understand speech. Am Sci 86(3):236–244
12. Wu S-L, Kingsbury E, Morgan N, Greenberg S (1998) Incorporating information from syllable-length time scales into automatic speech recognition. In: Proceedings of the 1998 IEEE international conference on acoustics, speech and signal processing, ICASSP'98 (Cat. No. 98CH36181), vol 2. IEEE, pp 721–724
13. Ganapathiraju A, Hamaker J, Picone J, Ordowski M, Doddington GR (2001) Syllable-based large vocabulary continuous speech recognition. IEEE Trans Speech Audio Process 9(4):358–366
14. Origlia A, Abete G, Cutugno F (2013) A dynamic tonal perception model for optimal pitch stylization. Comput Speech Lang 27(1):190–208
15. Origlia A, Cutugno F, Galatà V (2014) Continuous emotion recognition with phonetic syllables. Speech Commun 57:155–169
16. Wagner P, Origlia A, Avesani C, Christodoulides G, Cutugno F, d'Imperio M, Mancebo DE, Fivela BG, Lacharet A, Ludusan B et al (2015) Different parts of the same elephant: a roadmap to disentangle and connect different perspectives of prosodic prominence. In: International congress of phonetic sciences (ICPhS 2015). International Phonetic Association
17. Fujimura O (1994) Syllable timing computation in the c/d model. In: Third international conference on spoken language processing (ICLPS 1994), Yokohama, Japan, pp 519–522
18. Warren RM, Healy EW, Chalikia MH (1996) The vowel-sequence illusion: Intrasubject stability and intersubject agreement of syllabic forms. J Acoust Soc Am 100(4):2452–2461
19. Arnal LH, Poeppel D, Giraud A-L (2016) A neurophysiological perspective on speech processing in "the neurobiology of language". In: Neurobiology of language. Elsevier, Amsterdam, pp 463–478
20. Wu S-L, Kingsbury ED, Morgan N, Greenberg S (1998) Incorporating information from syllable-length time scales into automatic speech recognition. In: Proceedings of the 1998 IEEE international conference on acoustics, speech and signal processing, ICASSP'98 (Cat. No.98CH36181), vol 2, pp 721–7242
21. Huang X, Acero A, Hon H-W, Foreword By-Reddy R (2001) Spoken language processing: a guide to theory, algorithm, and system development. Prentice Hall PTR, Hoboken
22. Wang Y, Mohamed A, Le D, Liu C, Xiao A, Mahadeokar J, Huang H, Tjandra A, Zhang X, Zhang F et al (2020) Transformer-based acoustic modeling for hybrid speech recognition. In: ICASSP 2020–2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6874–6878
23. Batarseh FA, Yang R (2020) Data democracy: at the nexus of artificial intelligence, software development, and knowledge engineering. Academic Press, Cambridge
24. Awasthi P, George JJ (2020) A case for data democratization. In: Proceedings of the Americas conference on information systems (AMCIS)
25. Fujimura O (1975) Syllable as a unit of speech recognition. IEEE Trans Acoust Speech Signal Process 23(1):82–87
26. Ruske G, Schotola T (1978) An approach to speech recognition using syllabic decision units. In: ICASSP'78. IEEE international conference on acoustics, speech, and signal processing, vol 3. IEEE, pp 722–725
27. Kingsbury BE (1998) Perceptually inspired signal processing strategies for robust speech recognition in reverberant environments. University of California, Berkeley
28. Kingsbury BE, Morgan N, Greenberg S (1998) Robust speech recognition using the modulation spectrogram. Speech Commun 25(1–3):117–132
29. Wu S-L, Kingsbury B, Morgan N, Greenberg S (1998) Performance improvements through combining phone-and syllable-scale information in automatic speech recognition. In: ICSLP, vol 1, pp 160–163
30. Mogran N, Bourlard H, Hermansky H (2004) Automatic speech recognition: an auditory perspective. Springer, New York, pp 309–338. https://doi.org/10.1007/0-387-21575-1_6
31. Cutugno F, Coro G, Petrillo M (2005) Multigranular scale speech recognizers: technological and cognitive view. In: Congress of the Italian association for artificial intelligence. Springer, Berlin, pp 327–330
32. Panda SP, Nayak AK (2016) Automatic speech segmentation in syllable centric speech recognition system. Int J Speech Technol 19(1):9–18

33. Li J et al (2022) Recent advances in end-to-end automatic speech recognition. APSIPA Trans Signal Inf Process 11(1)
34. Karmakar P, Teng SW, Lu G (2021) Thank you for attention: a survey on attention-based artificial neural networks for automatic speech recognition. arXiv preprint arXiv:2102.07259
35. Graves A, Fernández S, Gomez F, Schmidhuber J (2006) Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd international conference on machine learning, pp 369–376
36. Graves A, Jaitly N (2014) Towards end-to-end speech recognition with recurrent neural networks. In: International conference on machine learning. PMLR, pp 1764–1772
37. Novoa J, Wuth J, Escudero JP, Fredes J, Mahu R, Yoma NB (2018) DNN-HMM based automatic speech recognition for HRI scenarios. In: 2018 13th ACM/IEEE international conference on human-robot interaction (HRI). IEEE, pp 150–159
38. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30
39. Yeh C-F, Mahadeokar J, Kalgaonkar K, Wang Y, Le D, Jain M, Schubert K, Fuegen C, Seltzer ML (2019) Transformer-transducer: end-to-end speech recognition with self-attention. arXiv preprint arXiv:1910.12977
40. Zhang S, Loweimi E, Bell P, Renals S (2021) On the usefulness of self-attention for automatic speech recognition with transformers. In: 2021 IEEE Spoken language technology workshop (SLT). IEEE, pp 89–96
41. Humphreys GW, Sui J (2016) Attentional control and the self: the self-attention network (SAN). Cogn Neurosci 7(1–4):5–17
42. Clark K, Khandelwal U, Levy O, Manning CD (2019) What does BERT look at? An analysis of BERT's attention. In: Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP. Association for Computational Linguistics, Florence, pp 276–286. https://doi.org/10.18653/v1/W19-4828
43. Subakan C, Ravanelli M, Cornell S, Bronzi M, Zhong J (2021) Attention is all you need in speech separation. In: ICASSP 2021–2021 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 21–25
44. Sun X, Yang D, Li X, Zhang T, Meng Y, Qiu H, Wang G, Hovy E, Li J (2021) Interpreting deep learning models in natural language processing: a review. arXiv preprint arXiv:2110.10470
45. Gulati A, Qin J, Chiu C-C, Parmar N, Zhang Y, Yu J, Han W, Wang S, Zhang Z, Wu Y et al (2020) Conformer: convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100
46. Graves A (2012) Sequence transduction with recurrent neural networks. arXiv preprint arXiv:1211.3711
47. Bengio Y, Goodfellow I, Courville A (2017) Deep learning, vol 1. MIT Press, Cambridge
48. Gunning D (2017) Explainable artificial intelligence (XAI). Defense advanced research projects agency (DARPA). nd Web 2(2):1
49. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang G-Z (2019) XAI-explainable artificial intelligence. Sci Robotics 4(37):7120
50. Xu F, Uszkoreit H, Du Y, Fan W, Zhao D, Zhu J (2019) Explainable AI: a brief survey on history, research areas, approaches and challenges. In: Natural language processing and chinese computing: 8th CCF international conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, proceedings, Part II 8. Springer, Berlin, pp 563–574
51. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R et al (2020) Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82–115
52. Dargan S, Kumar M, Ayyagari MR, Kumar G (2020) A survey of deep learning and its applications: a new paradigm to machine learning. Arch Comput Methods Eng 27:1071–1092
53. Smys S, Chen JIZ, Shakya S (2020) Survey on neural network architectures with deep learning. J Soft Comput Paradig (JSCP) 2(03):186–194
54. Ahmed I, Jeon G, Piccialli F (2022) From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where. IEEE Trans Ind Inform 18(8):5031–5042
55. Moshayedi AJ, Roy AS, Kolahdooz A, Shuxin Y (2022) Deep learning application pros and cons over algorithm deep learning application pros and cons over algorithm. EAI Endorsed Trans AI Robotics 1(1)
56. Dwivedi R, Dave D, Naik H, Singhal S, Omer R, Patel P, Qian B, Wen Z, Shah T, Morgan G et al (2023) Explainable AI (XAI): core ideas, techniques, and solutions. ACM Comput Surv 55(9):1–33
57. Shlezinger N, Whang J, Eldar YC, Dimakis AG (2023) Model-based deep learning. Proc IEEE
58. Piccialli F, Di Somma V, Giampaolo F, Cuomo S, Fortino G (2021) A survey on deep learning in medicine: why, how and when? Inf Fusion 66:111–137
59. Basak S, Agrawal H, Jena S, Gite S, Bachute M, Pradhan B, Assiri A (2023) Challenges and limitations in speech recognition technology: a critical review of speech signal processing algorithms, tools and systems. CMES Comput Model Eng Sci 135(2):1053–1089
60. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H (2019) Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov 9(4):1312
61. Kailkhura B, Gallagher B, Kim S, Hiszpanski A, Han T (2019) Reliable and explainable machine-learning methods for accelerated material discovery. npj Comput Mater 5(1):1–9
62. Bai X, Wang X, Liu X, Liu Q, Song J, Sebe N, Kim B (2021) Explainable deep learning for efficient and robust pattern recognition: a survey of recent developments. Pattern Recogn 120:108102
63. Angelov P, Soares E (2020) Towards explainable deep neural networks (xDNN). Neural Netw 130:185–194
64. Mziou-Sallami M, Khalsi R, Smati I, Mhiri S, Ghorbel F (2023) DeepGCSS: a robust and explainable contour classifier providing generalized curvature scale space features. Neural Comput Appl 1–12
65. Sahyoun A, Shehata S (2023) Aradiawer: an explainable metric for dialectical Arabic ASR. In: Proceedings of the second workshop on NLP applications to field linguistics, pp 64–73
66. Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H (2015) Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579
67. Alain G, Bengio Y (2016) Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644
68. Belinkov Y (2022) Probing classifiers: promises, shortcomings, and advances. Comput Linguist 48(1):207–219
69. Apicella A, Isgrò F, Prevete R, Tamburrini G (2019) Contrastive explanations to classification systems using sparse dictionaries. In: International conference on image analysis and processing. Springer, Berlin, pp 207–218
70. Amodei D, Ananthanarayanan S, Anubhai R, Bai J, Battenberg E, Case C, Casper J, Catanzaro B, Cheng Q, Chen G et al (2016) Deep speech 2: end-to-end speech recognition in English and mandarin. In: International conference on machine learning. PMLR, pp 173–182
71. Belinkov Y, Ali A, Glass J (2019) Analyzing phonetic and graphemic representations in end-to-end automatic speech recognition. arXiv preprint arXiv:1907.04224
72. Viglino T, Motlicek P, Cernak M (2019) End-to-end accented speech recognition. In: Interspeech, pp 2140–2144

73. Prasad A, Jyothi P (2020) How accents confound: probing for accent information in end-to-end speech recognition systems. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 3739–3753
74. Baevski A, Zhou Y, Mohamed A, Auli M (2020) wav2vec 2.0: A framework for self-supervised learning of speech representations. Adv Neural Inf Process Syst 33:12449–12460
75. Ma D, Ryant N, Liberman M (2021) Probing acoustic representations for phonetic properties. In: ICASSP 2021–2021 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 311–315
76. Pasad A, Chou J-C, Livescu K (2021) Layer-wise analysis of a self-supervised speech representation model. In: 2021 IEEE automatic speech recognition and understanding workshop (ASRU). IEEE, pp 914–921
77. Li C-Y, Yuan P-C, Lee H-Y (2020) What does a network layer hear? Analyzing hidden representations of end-to-end ASR through speech synthesis. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6434–6438
78. Coro G, Cutugno F, Caropreso F (2007) Speech recognition with factorial-hmm syllabic acoustic models. In: INTERSPEECH, pp 870–873
79. Wu Y, Zhang R, Rudnicky A (2007) Data selection for speech recognition. In: 2007 IEEE workshop on automatic speech recognition & understanding (ASRU). IEEE, pp 562–565
80. Namdari A, Li Z (2019) A review of entropy measures for uncertainty quantification of stochastic processes. Adv Mech Eng 11(6):1687814019857350
81. Xueli L, Hui D, Boling X (2005) Entropy-based initial/final segmentation for Chinese whiskered speech. Acta Acustica 30(1):69–75
82. Kuo H-KJ, Gao Y (2006) Maximum entropy direct models for speech recognition. IEEE Trans Audio Speech Lang Process 14(3):873–881
83. Geudens A, Sandra D, Martensen H (2005) Rhyming words and onset-rime constituents: an inquiry into structural breaking points and emergent boundaries in the syllable. J Exp Child Psychol 92(4):366–387
84. Green CR, Diakite B (2008) Emergent syllable complexity in Colloquial Bamana. J West Afr Lang 35(1–2):45
85. Abate ST, Menzel W (2007) Automatic speech recognition for an under-resourced language-Amharic. In: Eighth annual conference of the international speech communication association
86. Nguyen HQ, Nocera P, Castelli E, Trinh VL (2008) Large vocabulary continuous speech recognition for Vietnamese, an under-resourced language. In: Spoken languages technologies for under-resourced languages
87. Seng S, Sam S, Le V-B, Bigi B, Besacier L (2008) Which unit for acoustic and language modeling for Khmer automatic speech recognition? In: International workshop on spoken languages technologies for under-resourced languages, pp 33–38 (2008)
88. Le V-B, Besacier L (2009) Automatic speech recognition for under-resourced languages: application to Vietnamese language. IEEE Trans Audio Speech Lang Process 17(8):1471–1482
89. Tachbelie MY, Abate ST, Besacier L (2014) Using different acoustic, lexical and language modeling units for ASR of an under-resourced language-Amharic. Speech Commun 56:181–194
90. Fendji JLKE, Tala DC, Yenke BO, Atemkeng M (2022) Automatic speech recognition using limited vocabulary: a survey. Appl Artif Intell 36(1):2095039
91. Nvidia (2022) Nvidia Nemo automatic speech recognition. Available at https://catalog.ngc.nvidia.com/orgs/nvidia/collections/nemo_asr
92. Nvidia (2022) Nvidia nemo ASR—small conformer-transducer english model, version 1.6.0. Available at https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_conformer_transducer_small
93. Nvidia (2022) Nvidia nemo ASR—medium conformer-transducer English model, version 1.6.0. Available at https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_conformer_transducer_medium
94. Nvidia (2022) Nvidia nemo ASR—large conformer-transducer English model, version 1.6.0. Available at https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_conformer_transducer_large
95. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
96. Wang Q, Guo Y, Yu L, Li P (2017) Earthquake prediction based on spatio-temporal data mining: an LSTM network approach. IEEE Trans Emerg Top Comput 8(1):148–158
97. Namdari A, Li ZS (2021) A multiscale entropy-based long short term memory model for lithium-ion battery prognostics. In: 2021 IEEE International conference on prognostics and health management (ICPHM). IEEE, pp 1–6
98. Namdari A, Samani MA, Durrani TS (2022) Lithium-ion battery prognostics through reinforcement learning based on entropy measures. Algorithms 15(11):393
99. Coro G, Bardelli S, Cuttano A, Scaramuzzo RT, Ciantelli M (2023) A self-training automatic infant-cry detector. Neural Comput Appl 35(11):8543–8559
100. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
101. Schettino L, Di Maro M, Origlia A, Cutugno F (2022) Phonetic variation and syllabic structures in Italian and German speech. Phonetik und Phonologie im deutschsprachigen Raum
102. Boersma P (2001) Praat, a system for doing phonetics by computer. Glot Int 5(9):341–345
103. Mary L, Yegnanarayana B (2008) Extraction and representation of prosodic features for language and speaker recognition. Speech Commun 50(10):782–796
104. Cutugno F, D'Anna L, Petrillo M, Zovato E (2002) APA: Towards an automatic tool for prosodic analysis. In: Speech Prosody 2002, international conference, pp 231–234
105. Bigi B, Meunier C, Nesterenko I, Bertrand R (2010) Automatic detection of syllable boundaries in spontaneous speech. In: 7th International conference on language resources and evaluation (LREC 2010), pp 3285–3292
106. Kumari R, Dev A, Kumar A (2022) An efficient syllable-based speech segmentation model using fuzzy and threshold-based boundary detection. Int J Comput Intell Appl 21(02):2250007
107. Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76(5):378
108. Boersma P et al (1993) Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: Proceedings of the institute of phonetic sciences, vol 17. Citeseer, pp 97–110
109. D'Anna L, Cutugno F (2003) Segmenting the speech chain into tone units: human behaviour vs automatic process. In: Proceedings of The XVth international congress of phonetic sciences (ICPhS), pp 1233–1236
110. Petrillo M, Cutugno F (2003) A syllable segmentation algorithm for English and Italian. In: Eighth European conference on speech communication and technology
111. Li Y, Anumanchipalli GK, Mohamed A, Lu J, Wu J, Chang EF (2022) Dissecting neural computations of the human auditory pathway using deep neural networks for speech. bioRxiv
112. Magnuson JS, You H, Luthra S, Li M, Nam H, Escabi M, Brown K, Allopenna PD, Theodore RM, Monto N et al (2020) Earshot: a minimal neural network model of incremental human speech recognition. Cogn Sci 44(4):12823

113. Millet J, Caucheteux C, Boubenec Y, Gramfort A, Dunbar E, Pallier C, King J-R et al (2022) Toward a realistic model of speech processing in the brain with self-supervised learning. Adv Neural Inf Process Syst 35:33428–33443
114. Mohamed A, Lee H-Y, Borgholt L, Havtorn JD, Edin J, Igel C, Kirchhoff K, Li S-W, Livescu K, Maaløe L et al (2022) Self-supervised speech representation learning: a review. IEEE J Sel Top Signal Process
115. Lippmann RP (1997) Speech recognition by machines and humans. Speech Commun 22(1):1–15
116. Bazzi I (2002) Modelling out-of-vocabulary words for robust speech recognition. PhD thesis, Massachusetts Institute of Technology
117. Liu Y, Fung P (2004) State-dependent phonetic tied mixtures with pronunciation modeling for spontaneous speech recognition. IEEE Trans Speech Audio Process 12(4):351–364
118. Chang X, Maekaku T, Fujita Y, Watanabe S (2022) End-to-end integration of speech recognition, speech enhancement, and self-supervised learning representation. arXiv preprint arXiv:2204.00540
119. Shih Y-J, Wang H-F, Chang H-J, Berry L, Lee H-Y, Harwath D (2023) Speechclip: integrating speech with pre-trained vision and language model. In: 2022 IEEE spoken language technology workshop (SLT). IEEE, pp 715–722
120. Schneider S, Baevski A, Collobert R, Auli M (2019) wav2vec: Unsupervised pre-training for speech recognition. arXiv preprint arXiv:1904.05862
121. Borsos Z, Marinier R, Vincent D, Kharitonov E, Pietquin O, Sharifi M, Roblek D, Teboul O, Grangier D, Tagliasacchi M et al (2023) Audiolm: a language modeling approach to audio generation. In: IEEE/ACM transactions on audio, speech, and language processing
122. Chang H-J, Yang S-W, Lee H-Y (2022) Distilhubert: speech representation learning by layer-wise distillation of hidden-unit BERT. In: ICASSP 2022–2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 7087–7091
123. Yoshimura T, Hayashi T, Takeda K, Watanabe S (2020) End-to-end automatic speech recognition integrated with CTC-based voice activity detection. In: ICASSP 2020—2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6999–7003
124. Masumura R, Ihori M, Takashima A, Tanaka T, Ashihara T (2020) End-to-end automatic speech recognition with deep mutual learning. In: 2020 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC). IEEE, pp 632–637
125. Rajasegaran J, Khan S, Hayat M, Khan FS, Shah M (2020) Self-supervised knowledge distillation for few-shot learning. arXiv preprint arXiv:2006.09785
126. Wang Y, Yao Q, Kwok JT, Ni LM (2020) Generalizing from a few examples: A survey on few-shot learning. ACM Comput Surv (CSUR) 53(3):1–34
127. Baevski A, Auli M, Mohamed A (2019) Effectiveness of self-supervised pre-training for speech recognition. arXiv preprint arXiv:1911.03912
128. Yan LJ, Ming LK, Yin OS, Poo LC (2021) Efficient-prototypicalnet with self knowledge distillation for few-shot learning. Neurocomputing 459(2021):327–337
129. Kim D, Kim G, Lee B, Ko H (2022) Prototypical knowledge distillation for noise robust keyword spotting. IEEE Signal Process Lett 29:2298–2302
130. Ludusan B, Origlia A, Cutugno F (2011) On the use of the rhythmogram for automatic syllabic prominence detection. In: Twelfth annual conference of the international speech communication association
131. Zlokarnik I (1995) Adding articulatory features to acoustic features for automatic speech recognition. J Acoust Soc Am 97(5-Supplement):3246–3246
132. Wu S-L, Shire ML, Greenberg S, Morgan N (1997) Integrating syllable boundary information into speech recognition. In: 1997 IEEE International conference on acoustics, speech, and signal processing, vol 2. IEEE, pp 987–990
133. King S, Taylor P, Frankel J, Richmond K (2000) Speech recognition via phonetically-featured syllables
134. Ramya R, Hegde RM, Murthy HA (2008) Incorporating acoustic feature diversity into the linguistic search space for syllable based speech recognition. In: 2008 16th European signal processing conference. IEEE, pp 1–5
135. Lee T, Liu Y, Huang P-W, Chien J-T, Lam WK, Yeung YT, Law TK, Lee KY, Kong AP-H, Law S-P (2016) Automatic speech recognition for acoustical analysis and assessment of Cantonese pathological voice and speech. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6475–6479
136. Wang X, Yao Z, Shi X, Xie L (2021) Cascade RNN-transducer: Syllable based streaming on-device mandarin speech recognition with a syllable-to-character converter. In: 2021 IEEE spoken language technology workshop (SLT). IEEE, pp 15–21
137. Vempada RR, Kumar BSA, Rao KS (2012) Characterization of infant cries using spectral and prosodic features. In: 2012 National conference on communications (NCC). IEEE, pp 1–5
138. Ji C, Xiao X, Basodi S, Pan Y (2019) Deep learning for asphyxiated infant cry classification based on acoustic features and weighted prosodic features. In: 2019 International conference on internet of things (iThings) and IEEE green computing and communications (GreenCom) and IEEE cyber, physical and social computing (CPSCom) and IEEE smart data (SmartData). IEEE, pp 1233–1240
139. Cohen R, Lavner Y (2012) Infant cry analysis and detection. In: 2012 IEEE 27th convention of electrical and electronics engineers in Israel. IEEE, pp 1–5
140. Liu L, Li Y, Kuo K (2018) Infant cry signal detection, pattern extraction and recognition. In: 2018 International conference on information and computer technologies (ICICT). IEEE, pp 159–163
141. Ji C, Mudiyanselage TB, Gao Y, Pan Y (2021) A review of infant cry analysis and classification. EURASIP J Audio Speech Music Process 2021(1):1–17
142. Olivier R, Raj B (2022) Recent improvements of ASR models in the face of adversarial attacks. arXiv preprint arXiv:2203.16536
143. Zhao Y, Calapodescu I (2022) Multimodal robustness for neural machine translation. In: Proceedings of the 2022 conference on empirical methods in natural language processing, pp 8505–8516
144. Xue L, Gao M, Chen Z, Xiong C, Xu R (2023) Robustness evaluation of transformer-based form field extractors via form attacks. In: International conference on document analysis and recognition. Springer, Berlin, pp 167–184
145. Wu Y, Xu X, Walker PR, Liu J, Saxena N, Chen Y, Yu J (2021) HVAC Evading classifier-based defenses in hidden voice attacks. In: Proceedings of the 2021 ACM Asia conference on computer and communications security, pp 82–94
146. Zhang Z, Yang E, Fang S (2021) Commandergabble: a universal attack against ASR systems leveraging fast speech. In: Annual computer security applications conference, pp 720–731
147. Assante M, Candela L, Castelli D, Cirillo R, Coro G, Frosini L, Lelii L, Mangiacrapa F, Pagano P, Panichi G et al (2019) Enacting open science by d4science. Futur Gener Comput Syst 101:555–563

148. Coro G, Panichi G, Scarponi P, Pagano P (2017) Cloud computing in a distributed e-infrastructure using the web processing service standard. Concurr Comput Pract Exp 29(18):4219
149. Coro G, Candela L, Pagano P, Italiano A, Liccardo L (2015) Parallelizing the execution of native data mining algorithms for computational biology. Concurr Comput Pract Exp 27(17):4630–4644