On incrementing interpretability of machine learning models from the foundations: a study on syllabic speech units

Vincenzo Norman Vitale[1,2], Loredana Schettino[1] and Francesco Cutugno[1,2]

[1]Interdepartmental Research Center Urban/Eco, University of Naples Federico II, Italy
[2]DIETI - University of Naples Federico II, Italy, Italy

Abstract

English. Modern ASR systems generally encode information by employing representations that favour performance indicators such as Word Error Rate (WER), making the interpretation of results and the diagnosis of any error extremely difficult if not impossible. In particular, within the context of end-to-end ASR systems, studies have been devoted to investigating the degrees of explainability of such systems by considering the use of different sets of linguistic features. This work explores the potential of different machine learning algorithms by considering features extracted from syllabic units of analysis and highlights that relying on syllabic Mel-Frequency Cepstral Coefficients increases the interpretability of complex techniques. In fact, the latter currently extract basic units in ways that are highly skewed toward operational convenience. The proposed method would reduce the need for computational resources both in training and in the inference phases, which results in economical and less time-consuming processes.

1. Introduction

[1] The advent of Deep Neural Networks (DNN) enabled modern ASR systems and, more in general, Natural Language Processing (NLP) systems to perform at their best when fed with enough training data and supplied with sufficient computational resources. The recent tendency is to focus efforts on incrementing performance indicators like Word Error Rate (WER), making DNN models behind the scenes increasingly complex and larger, with the effect of a dramatic reduction in their interpretability and an increase in the number of parameters considered and therefore in the required computation effort [1, 2]. As an example, state-of-the-art End-to-End (E2E) ASR systems [3, 4, 5, 6] employ self-supervised learning techniques to determine, based on huge amounts of unlabelled data, the best representation of the speech signal based on fixed-length units, which results in adaptable systems. In the same way, Big Language Models (Big LM) employ advanced encoding techniques, like those based on Byte Pair Encoding (BPE)[7, 8, 9], to encode sub-word units reducing their impact on memory and thus allowing for the creation of bigger models with millions or even billions of parameters aimed at catching a wider range of natural language nuances. On the one hand, these techniques definitely improve systems' performances and capabilities. On the other hand, they also reduce models' interpretability from their foundations, which not only makes them increasingly similar to black boxes but also augments their need for computational resources. Wav2Vec2 authors [3] suggest that "switching to a seq2seq architecture and a word piece vocabulary" would result in performance gains. In line with this, the employment of larger and linguistically motivated units, like syllables, could bring several advantages. Firstly, it would improve performance in terms of WER and computational resources required to train and operate these systems. Secondly, it would increment the system's interpretability, allowing domain experts (i.e. linguists, especially phoneticians) to dive deep into error analysis, which means favoring interpretable rather than computationally efficient but poorly understandable inputs. The main contributions of this study are:

- the proposal of an interpretable approach to speech-oriented feature extraction based on syllable;
- a comparison of various classification techniques with different interpretability grades.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
vincenzonorman.vitale@unina.it (V. N. Vitale); loredana.schettino@unina.it (L. Schettino); cutugno@unina.it (F. Cutugno)
0000-0002-0365-8575 (V. N. Vitale); 0000-0002-3788-3754 (L. Schettino); 0000-0001-9457-6243 (F. Cutugno)

[1]This article results from the collaboration among the authors. However, for academic purposes, Norman Vincenzo Vitale is responsible for sections 1 and 2.1 and 3.3, Loredana Schettino for sections 2.2, 3.1 and 3.2 for section 4. All the authors are responsible for sections 4 and 5.

2. Related Work

2.1. Explaining Modern ASR

Among the drawbacks of modern Deep Neural Networks (DNN) based systems, the most frequently cited are their poor interpretability [10], the lack of sufficient training corpora [11] and the high demand for computational resources [12, 13, 14]. In recent years, studies have been devoted to the investigation of the degrees of explainability of ASR and, more in general, of speech-related systems based on DNN. Some of these works aim to interpret the internal model dynamics and overall 'behaviour' through model-output backtracing or simulations via explainability methods [15, 16, 17]. In some other cases, 'probing'techniques have been employed to investigate what's encoded in DNN layers at different 'depths' [18, 19] by introducing probes aimed at catching intermediate internal representations to be used for various tasks (e.g., regression or classification). Through classification and measurements, some probing studies have analysed how the accent of pronunciation in different English varieties influences the performance of DeepSpeech2 [20, 21]. These studies employed linguistic-related features and highlighted how the contextual phonetic information contained in intermediate representations influenced the classification. A further study used probing to investigate the multi-temporal modelling of phonetic information in the Wav2Vec2 ASR [3, 22]. Some authors have also proposed a spectrogram-like representation of emissions that could be used for speaker identification and speech synthesis [23].

However, although some studies did consider linguistic features to explain the behaviour of existing models, these were related to isolated fixed-length segments and, to the best of our knowledge, did not take into account larger linguistically meaningful units (i.e. syllables).

2.2. Syllabic unit structure

The notion of *syllable* is quite well-known in linguistic studies. However, its definition has been much debated since it consists of dynamic and complex structures that can be analysed from different perspectives, i.e., phonological or phonetics, and involve various aspects, like articulatory gestures coordination, intensity modulation [24]. Acoustically, a syllable is described as being characterized by an intensity peak in the speech signal surrounded by less intense aggregated sounds.

The essential element of a syllabic unit is the sonority peak, the *nucleus*, which usually consists of vowel sounds. The nucleus can be accompanied by aggregated consonant sounds at the beginning of the unit preceding the nucleus, the *onset*, or at the end following it, the *coda*. Different languages allow syllabic combinations of vocalic (V) and consonantal (C) sounds with different degrees of complexity. However, in many languages, CV is described as the most common structure and the most resistant to phonetic variation and the related reduction phenomena [25, 26, 27].

While disagreements have mainly concerned determining boundaries between different units, the alternation of different units can be grounded on the principles of sonority scale and onset maximization [28]. Thus, the syllable can be described as a sequence of speech sounds where the onset of the sequence is less intense than the preceding coda.

Based on the structural integrity of the syllable, evidence has been provided that the syllable rather than phonetic segments can represent a relevant basic unit of speech production and perception [29, 30]. In fact, [26] shows that the observable variation in connected speech is more systematic at the level of the syllable than at one of the phonetic segments.

3. Material and Methods

To achieve our defined goal, we considered an input consisting of syllables from datasets manually annotated by domain experts and evaluated the performances achieved by different classification methods when relying on distinct sets of syllable-based features.

3.1. Corpus and annotation

This study is based on the Italian and Spanish datasets of the Nocando corpus [31] which consists of spoken narrative texts produced by 11 Italian and 6 Spanish subjects.

The audio files and their transcriptions were processed using the WebMAUS Basic services [32] which provided automatic phonetic transcriptions. The latter were manually edited in Praat [33] and syllabified according to the principles of sonority sequencing and onset maximization [28]. Syllabic units were also annotated for their phonetic structural pattern (CV, CCV, CCVC, CVC, VC, V).

The Italian dataset consists of 940 syllables. As expected, the structural patterns are not evenly distributed, but the following distribution is observed: CV (65%), CVC (16%), CCV (9%), VC (3%), CCVC (3%), V (2%).

As for the Spanish dataset, it consists of 609 syllables. The occurring patterns do not differ much from the Italian ones: CV (65%), CVC (14%), CCV (7%), VC (5%), V (4%), CCVC (3%).

3.2. Syllable-based features

For our tasks, we assumed the hand-annotated syllables as base units. These were considered in two different

ways. At first, we look at syllables as a single piece of signal, which is how they have been traditionally considered and processed. Then, we consider them as a signal presumably made up of three components, namely the onset, the nucleus and the coda. The following four feature sets were considered.

- **OPSM** consisting of the GeMAPS [34] set from the OpenSmile toolkit [35]. It is composed of 62 features and provides information about the whole considered signal, namely the syllable.
- **MFCC** consists of 13 Mel Frequency Cepstrum Coefficients, which represent the most salient information for speech recognition tasks [36], extracted for each syllable part[2], i.e. onset, nucleus and coda.
- **Full** namely the concatenation OPSM and MFCC.
- **PCA** consists of the Principal Component Analysis [3] (with an explained variance 95%) of the Full set.

In order to avoid biases due to dimensionality, all the considered feature sets were normalized to achieve zero-mean and unitary variance [4].

3.3. The experimental protocol

This study consists of a classification task that concerns samples labelled with syllable patterns and aims at classifying them on the basis of the considered feature sets. In particular, we compared the following four techniques:

- The **K-Means**[3][38] a vector-quantization method which divides n objects in k clusters based on their mean distance.[5]
- **Hierarchical Agglomerative Clustering (HAC)**[3] [39] is a greedy technique which aims at grouping (or splitting) clusters based on a similarity measure. The final output is a clusters hierarchy which could be divided based on the number of desired clusters.[6]
- The **Support Vector Machine (SVM)**[3] [40] is a versatile algorithm used for classification and regression tasks, whose objective is to find a hyperplane in a multi-dimensional space that enables the classification of the considered data points.
- Lastly, we considered **Convolutional Neural Network (CNN)** [41, 42] as they represent state-of-the-art in speech processing tasks [3]. The considered CNN[7] consists of a Conv2d layer with the ReLu activation function, followed by a Feed-Forward with a SoftMax. In particular, we choose to compare two settings: the first one with a kernel size of 3x3; The second one considers a larger context with a size of 3x9.

In the first phase, we compared K-Means, HAC and SVM as, by their very nature, they are considered more interpretable than neural networks. On the one hand, K-Means and HAC allowed us to explore how sample grouping is affected when the numerousness of clusters is fixed or not, without external supervision. Then, SVM provides a robust and interpretable way of supervisingly evaluating how samples group when a model is set to learn a few interpretable parameters. Lastly, we evaluated the performance of a CNN on the MFCC feature set to compare it with the best-performing method among those from the previous phase, allowing us to compare how an interpretable yet powerful method performs against one of the fundamental building blocks of modern DNN. We compare performances through the micro averaged F1 score (Equation 1), which is particularly suitable for multi-classification tasks.

$$F1 = \frac{TP}{TP + \frac{1}{2} * (FP + FN)} \tag{1}$$

Given the fuzzy boundary between syllabic units and the high degree of variability within each syllable structure class, which not only concerns the presence or absence of segments but also their phonetic specification, the described techniques are applied considering the pooled types of syllable samples.

4. Results

Figure 1 reports the F1-score achieved by K-Means, HAC and SVM over all the considered feature sets. The SVM classifier outperforms both clustering methods on any feature set. However, this was not our primary goal. Note how, for any of the considered methods, the performance difference between the MFCC set and the PCA one is rather small.

For the SVM, results reported are referred to the optimal configuration, which has been found through a grid search on C (between 0.5 and 10 with a step of 0.5), gamma (within 0.01, 0.001 and 0.0001) and kernel type (within rbf, polynomial and sigmoid). The train, validation and dev set were respectively 60/20/20 of the original

[2] MFCC components were extracted through the Librosa library at version 0.9.2.

[3] Defined in Scikit-learn [37] version 1.1.3.

[4] Normalization was achieved through the Scikit-learn (version 1.1.3) StandardScaler.

[5] K-Means parameters: k=6, tolerance to declare convergence=1e-4, initialization through the k-means++ method, random state=42, algorithm=LLoyd's EM algorithm.

[6] HAC parameters: clusters=6, metric=euclidean, linkage=ward

[7] Implemented with Pytorch 1.13 and Pytorch-lightning 1.8.3

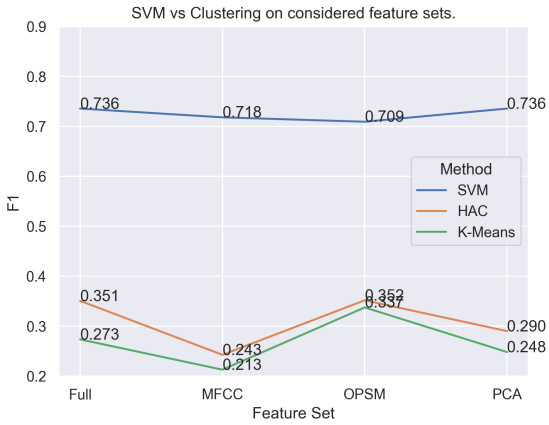

Figure 1: F-scores of SVM, K-Means and HAC per all the considered feature sets.

dataset. Data splits were balanced on the combination of the pattern (i.e. the label) and language.

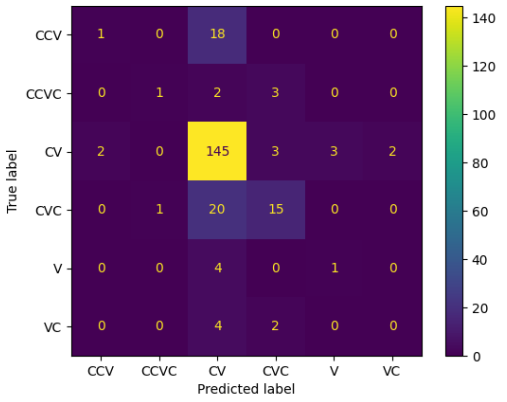

Figure 2: Confusion matrix of the SVM classifier on the MFCC feature set

The confusion matrix that reports the output of the SVM classifier operating on the basis of MFCC (Fig. 2), highlights that better performances concern the CV structure, which is also the more frequent in the data. As for the other structures, misclassification cases mostly concern their identification as CV, which reveals that when considering syllabic units actually occurring in the speech signal a particularly high similarity emerges between more complex patterns, i.e. CCV and CVC, and the CV pattern.

Lastly, Fig. 3 reports the comparison between the SVM classifier and the considered CNN configurations on the MFCC feature set. Still, the SVM performed better than both CNN-based configurations.

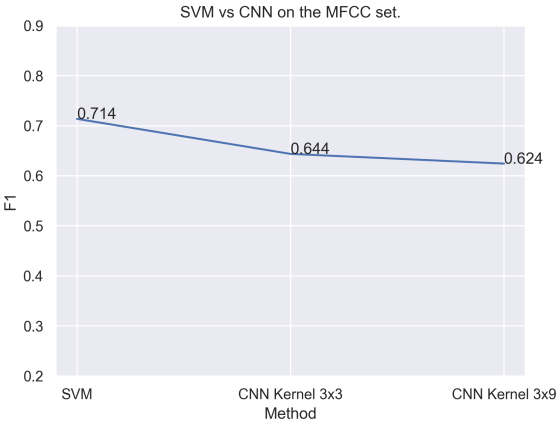

Figure 3: Comparison between SVM and CNN settings on the MFCC feature set.

5. Discussion and Conclusions

In this study, we evaluated the use of phonetic syllables as basic units for speech-related tasks aimed at preserving and, if possible, incrementing the interpretability of different learning techniques. We employed four different feature sets extracted upon the assumption of the phonetic syllable as a fundamental unit, considering different points of view: the more theoretically informed MFCC-based that is strongly tied to the signal; the more analytic OPSM based on Opensmile statistical analysis; the most extended which is a combination of both; the most computationally efficient based on the PCA analysis. Then, upon these feature sets, we evaluated the performance of three well-known machine learning techniques known for being highly interpretable. Finally, we compared the best-performing model, namely the SVM, with a convolutional network on the MFCC set, obtaining comparable performances. Our preliminary results highlight that a set of features aimed at keeping things interpretable, namely the MFCC, lets different methods achieve performances that are comparable to those of richer (Full), analytic (OPSM) or computationally optimized (PCA) sets, which do not retain the same interpretability grade. These findings corroborate the idea

that training speech-oriented learning models on larger and linguistically meaningful units could increase the capacity of domain experts and software/ml engineers to diagnose system failures and, at the same time, help reduce the effort and computational resources needed for signal preprocessing. Ongoing analyses involve the enlargement of the annotated datasets to improve the results of further classification trials. In Appendix we reported some preliminary results of a classification trial on an extended dataset, about 26 minutes of hand-annotated speech, consisting of 3589 phonetic syllables. In future works, we plan to extend this kind of study to recent architectures like Squeezeformer[43] or CNN-BLSTM[44].

References

[1] Y. Belinkov, A. Ali, J. Glass, Analyzing phonetic and graphemic representations in end-to-end automatic speech recognition, arXiv preprint arXiv:1907.04224 (2019).

[2] D. V. Carvalho, E. M. Pereira, J. S. Cardoso, Machine learning interpretability: A survey on methods and metrics, Electronics 8 (2019) 832.

[3] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems 33 (2020) 12449–12460.

[4] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, Hubert: Self-supervised speech representation learning by masked prediction of hidden units, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 3451–3460.

[5] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, M. Auli, Data2vec: A general framework for self-supervised learning in speech, vision and language, in: International Conference on Machine Learning, PMLR, 2022, pp. 1298–1312.

[6] Z. Zhang, L. Zhou, J. Ao, S. Liu, L. Dai, J. Li, F. Wei, Speechut: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 1663–1676.

[7] R. Sennrich, B. Haddow, A. Birch, Neural machine translation of rare words with subword units, arXiv preprint arXiv:1508.07909 (2015).

[8] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, The Journal of Machine Learning Research 21 (2020) 5485–5551.

[10] Y. Zhang, P. Tiňo, A. Leonardis, K. Tang, A survey on neural network interpretability, IEEE Transactions on Emerging Topics in Computational Intelligence 5 (2021) 726–742.

[11] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., A survey of large language models, arXiv preprint arXiv:2303.18223 (2023).

[12] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in nlp, arXiv preprint arXiv:1906.02243 (2019).

[13] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, J. Dean, Carbon emissions and large neural network training, arXiv preprint arXiv:2104.10350 (2021).

[14] D. Patterson, J. Gonzalez, U. Hölzle, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. R. So, M. Texier, J. Dean, The carbon footprint of machine learning training will plateau, then shrink, Computer 55 (2022) 18–28.

[15] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9 (2019) e1312.

[16] B. Kailkhura, B. Gallagher, S. Kim, A. Hiszpanski, T. Han, Reliable and explainable machine-learning methods for accelerated material discovery, npj Computational Materials 5 (2019) 1–9.

[17] P. Angelov, E. Soares, Towards explainable deep neural networks (xdnn), Neural Networks 130 (2020) 185–194.

[18] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, H. Lipson, Understanding neural networks through deep visualization, arXiv preprint arXiv:1506.06579 (2015).

[19] G. Alain, Y. Bengio, Understanding intermediate layers using linear classifier probes, arXiv preprint arXiv:1610.01644 (2016).

[20] T. Viglino, P. Motlicek, M. Cernak, End-to-end accented speech recognition., in: Interspeech, 2019, pp. 2140–2144.

[21] A. Prasad, P. Jyothi, 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, 2020, pp. 3739–3753.

[22] D. Ma, N. Ryant, M. Liberman, Probing acoustic representations for phonetic properties, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 311–315.

[23] C.-Y. Li, P.-C. Yuan, H.-Y. Lee, What does a net-

work 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, 2020, pp. 6434–6438.

[24] J. Laver, L. John, Principles of phonetics, Cambridge university press, 1994.

[25] P. Maturi, I suoni delle lingue, i suoni dell'italiano: nuova introduzione alla fonetica, Bologna: il Mulino, 2014.

[26] S. Greenberg, Speaking in shorthand–a syllable-centric perspective for understanding pronunciation variation, Speech Communication 29 (1999) 159–176.

[27] L. Schettino, V. N. Vitale, F. Cutugno, Syllabic reduction in Italian connected speech: towards the integration of linguistic and computational approaches, in: Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS 2023), 2023, pp. 2015–2019.

[28] M. Nespor, Fonologia, Bologna: Il Mulino, 1993.

[29] F. Albano Leoni, The boundaries of the syllable, in: D. Russo (Ed.), The Notion of Syllable across History, Theories and Analysis, Cambridge Scholars Publishing, 2016.

[30] F. Cangemi, O. Niebuhr, Rethinking reduction and canonical forms, Rethinking reduction (2018) 277–302.

[31] L. Brunetti, S. Bott, J. Costa, E. Vallduví, A multilingual annotated corpus for the study of information structure 1, in: Grammatik und Korpora 2009. Dritte internationale Konferenz, Mannheim, 22.-24.09. 2009. Grammar & corpora 2009, 2011.

[32] T. Kisler, U. Reichel, F. Schiel, Multilingual processing of speech via web services, Computer Speech & Language 45 (2017) 326–347.

[33] P. Boersma, D. Weenink, Praat: doing phonetics by computer [computer program]. version 5.3. 51, Online: http://www. praat. org/retrieved, last viewed on 12 (1999-2022).

[34] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, et al., The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing, IEEE transactions on affective computing 7 (2015) 190–202.

[35] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 1459–1462.

[36] X. Huang, A. Acero, H.-W. Hon, R. Reddy, Spoken language processing: A guide to theory, algorithm, and system development, Prentice hall PTR, 2001.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[38] J. A. Hartigan, M. A. Wong, Algorithm as 136: A k-means clustering algorithm, Journal of the royal statistical society. series c (applied statistics) 28 (1979) 100–108.

[39] F. Murtagh, A survey of recent advances in hierarchical clustering algorithms, The computer journal 26 (1983) 354–359.

[40] W. S. Noble, What is a support vector machine?, Nature biotechnology 24 (2006) 1565–1567.

[41] K. O'Shea, R. Nash, An introduction to convolutional neural networks, arXiv preprint arXiv:1511.08458 (2015).

[42] S. Albawi, T. A. Mohammed, S. Al-Zawi, Understanding of a convolutional neural network, in: 2017 international conference on engineering and technology (ICET), Ieee, 2017, pp. 1–6.

[43] S. Kim, A. Gholami, A. Shaw, N. Lee, K. Mangalam, J. Malik, M. W. Mahoney, K. Keutzer, Squeezeformer: An efficient transformer for automatic speech recognition, Advances in Neural Information Processing Systems 35 (2022) 9361–9373.

[44] D. Wang, X. Wang, S. Lv, End-to-end mandarin speech recognition combining cnn and blstm, Symmetry 11 (2019) 644.

A. Further results

In Figures 5 and 4, we reported the confusion matrix of the best-performing SVM classifier on MFCC features, in the normalized and non-normalized versions respectively. These preliminary results were obtained on an extended version of the dataset which is currently under further analysis.

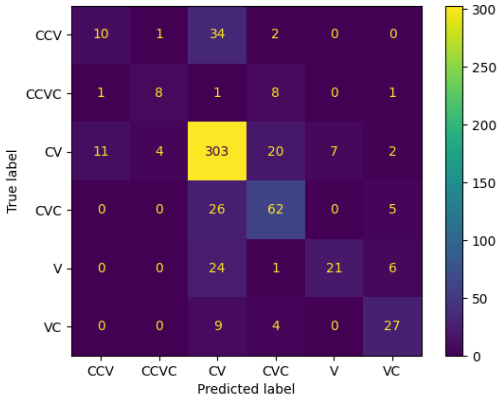

Figure 4: Confusion matrix of the SVM classifier on the MFCC feature set, based on a new version of the dataset incremented by 20 minutes of annotated speech.

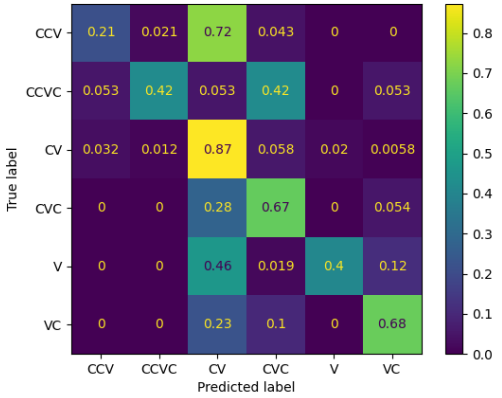

Figure 5: Normalized Confusion matrix of the SVM classifier on the MFCC feature set, based on a new version of the dataset incremented by 20 minutes of annotated speech.