Computational Paralinguistics: The Importance of Audio Analysis and Feature Extraction Methodologies

Mercedes Kiss-Vetráb

Booklet of the PhD Thesis

Supervisor:
Dr. Gábor Gosztolya

Doctoral School of Computer Science
University of Szeged

Department of Computer Algorithms and Artificial Intelligence

Szeged
2025

1 Introduction

Human communication extends far beyond spoken words. In every human interaction, numerous non-verbal cues accompany verbal expressions. For instance, tone of voice, temporal dynamics, and subtle intonations convey emotions, intent, speaker state, and meaning. The computational analysis of such paralinguistic information represents a rapidly advancing field in speech technology [23].

Despite rapid advances in machine learning and speech processing, the field of computational paralinguistics still lacks unified methodological frameworks. One of the significant challenges is dealing with small corpora or limited amounts of labelled data. One reason for this is that each use case usually requires specific recording protocols and annotations. Another crucial technical challenge is getting a fixed-sized feature vector from variable-length speech segments. In computational paralinguistics, we aim to assign a single label output (classification or regression) to an audio recording (utterance) of varying length. The ability to automatically infer emotions, speaker identities, and other paralinguistic attributes from speech signals has the potential to support a wide range of domains, such as healthcare, customer service, education, and entertainment. This motivation directly highlight the development of task-independent solutions, which aim to capture general speech characteristics across various paralinguistic tasks. It is a crucial step towards addressing the technical challenges in this field while simultaneously advancing its real-world applications. [24, 25].

The PhD thesis addresses the previously mentioned challenges by investigating feature extraction methodologies and architectural design choices. Through a comprehensive analysis, it aims to establish global guidelines for specific algorithmic decisions. The research include three interconnected streams: Traditional Feature Extraction Methods (Bag-of-Audio-Words (BoAW)), Hybrid Approaches (HMM/DNN Integration), and Deep Learning Techniques (Sequence-to-Sequence Autoencoder (Seq. Autoencoder) and Wav-to-Vec 2.0 Neural Network (Wav2Vec 2.0)). These experiments highlights the need for robust, generalisable solutions and provides practical guidelines across diverse paralinguistic applications.

The dissertation consists of 4 major parts. The first chapter presents the most important fundamental concepts, including technical challenges, databases and evaluation methodologies employed throughout the experiments. The second chapter presents results with traditional feature extraction methodologies, focusing on the BoAW technique. This chapter provides an understanding of how traditional machine learning approaches handle paralinguistic tasks and why parameter optimisation is crucial for achieving robust performance. The third chapter presents hybrid methodologies that combine traditional statistical methods with modern deep learning approaches. It provides a detailed analysis of the HMM/DNN hybrid system. It explains why selecting the proper audio preprocessing and aggregation strategy is fundamental to achieving optimal performance. It is bridging the gap between traditional and State-of-the-Art approaches. The fourth chapter presents SotA Deep Neural Network models. It describes how modern deep learning approaches, specifically the Seq. Autoencoder and Wav2Vec 2.0 models, can be optimised for paralinguistic tasks. The chapter includes experiments in audio preprocessing optimisation and advanced aggregation strategies. These experiments collectively address the critical need for robust, generalisable solutions in computational paralinguistics while providing practical guidelines across diverse paralinguistic applications.

2 Bag-of-Audio-Words as a Traditional Feature Extraction Method

Traditional machine learning models have long served as the backbone of computational paralinguistics. These techniques are especially vital in scenarios where the amount of available data or computational resources are limited. This makes these methods a good choice for rapid prototyping, testing, and deployment in real-world and low-resource environments. Additionally, traditional methods provide reliable benchmarks for evaluating new techniques, ensuring that innovations are significantly advance the field [23, 24].

In Chapter 3. of the PhD thesis, a comprehensive study of the Bag-of-Audio-Words method is presented. The effectiveness of this technique depends on the careful optimisation of architectural decisions made during its implementation. Figure 1 provides a step-by-step introduction of the technique. It is clustering frame-level features into ”codebooks” and based on them, creating histogram representations as fixed-sized utterance-level feature vectors. It provides compact and representative features of audio signals. The adjustable parameters of this methodology affect the quality of the extracted utterance-level features and the efficiency of the final classification or regression as well. The chapter’s main contributions include parameter optimisation strategies, demonstration of corpus independence capabilities, and methods for handling the stochastic nature of the BoAW algorithm.

2.1 Parameter Optimisation

This section details systematic experiments to optimise parameters of the BoAW method. This investigation addresses

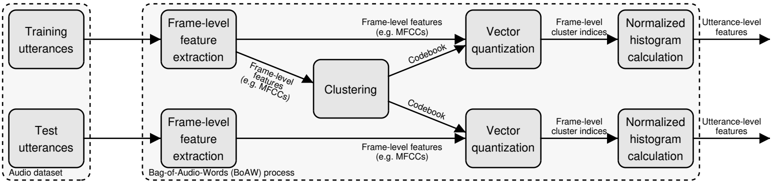
Training utterances
Test utterances
Audio dataset
Frame-level feature extraction
Frame-level feature extraction
Frame-level features (e.g. MFCCs)
Frame-level features (e.g. MFCCs)
Frame-level features (e.g. MFCCs)
Clustering
Codebook
Codebook
Vector quantization
Vector quantization
Frame-level cluster indices
Frame-level cluster indices
Normalized histogram calculation
Normalized histogram calculation
Utterance-level features
Utterance-level features
Bag-of-Audio-Words (BoAW) process

Figure 1: *Workflow for the Bag-of-Audio-Words technique.*

the following critical research question: *How do individual parameter choices affect classification/regression performance, and can we establish optimal parameter ranges?*

In this research, the experiments are conducted on an emotion detection task. Built upon the BoAW features, a Support Vector Machine (SVM) classification model was trained and evaluated on a Hungarian emotion database [27]. Parameters got iteratively optimised, taking into account the findings from each iteration. The following design choices were tested: preprocessing method (normalisation and standardisation), quantisation neighbour number (5 and 10), clustering method (k-means and k-means++), resampling method (upsampling), the effect of Δvalues, and the codebook size N (32, 64, 128, 256, 512, $1\,024$, $2\,048$, $4\,096$, $8\,192$).

Results indicated that precise tuning of multiple parameters was necessary to achieve optimal efficiency. We are provided clear recommendations on parameter configurations:

- Transform the input dataset to the same scale by normalisation or standardisation. Preprocessing is always a good choice.

- For greater generalisation ability, it is worth including more neighbours in the quantising step, such as 5 or 10.

- It is worth choosing the size of the codebook from a medium to large range (between 128 and $4\,096$). Try to keep the codebook size low to get better generalisation. The connection between increasing codeword quantities and decreasing model performance suggests that the larger the codebook size we choose, the greater the chance of overfitting.

- Clustering with the k-means or with the k-means++ algorithms could be equally good.
- By balancing the frequency of classes seen during learning, we can improve our generalisation ability. Upsampling can help to achieve it.
- With the Δvalues, we can reduce the number of necessary codewords to a moderate size, and the training trends are much more predictable.

2.2 Corpus Independence

In the previous part, the experiments were conducted in a corpus-dependent environment. The initial clustering step is typically corpus-dependent and performed on the training set of the database used. However, this approach has limitations, including the need to create new codewords for each dataset, which results in increased computational time and potential overfitting. The next experiment addresses the following critical research question: *To what extent is the BoAW approach dependent on corpus-specific characteristics, and can codebooks be transferred across different datasets?*

In this experiment, 3 databases were used: the Hungarian emotion database [27], the EmoDB [14], and the BEA [22] speech database. The classification was performed with an SVM model. In the first part, codebooks were created on *EmoDB*. With them, a BoAW representation was constructed for the Hungarian emotion database and used for classification. In the second part, the question arose whether the type of the database used to create the codebook can affect the success of the classification. Several experiments examined predefined codebooks to determine which setup is better: using databases created for similar purposes in a different language, or using databases for a similar language but with a different purpose. In the second part, new codebooks were prepared on different subsets (1-hour, 2-hour, 5-hour, and 10-hour) of the BEA database.

The results indicate a similar amount of performance improvement across all experiments, suggesting that the BoAW codebooks have practical corpus independence. Based on the results, it can be stated that predefined codebooks can be successfully used to extract BoAW feature representations for other databases. There was no significant difference between databases created for similar purposes, but in different languages, or for similar languages but different purpose.

2.3 Ensembling Strategy

The BoAW method contains non-consistent clustering, including random initialisation, so it can yield slightly different results for each re-run, even when using the same settings. This experiment addresses the following critical research question: *How does the inherent stochastic nature of the clustering process affect result reliability, and what strategies can mitigate this variability?*

In this experiment, the Sleepiness database [19] was used. Therefore, an Support Vector Regression Machine model was trained for predictions in a scale of 1 to 10. To test the robustness of the BoAW algorithm, we ran the feature extraction process 10 times with 10 different random start-up seeds. The following seeds were used: 1964, 423, 1355, 86, 1052, 1549, 139, 731, 951. We tried to handle the stochastic behaviour with 3 different ensembling methodologies:

- Model performance ensembling - Average models: From the feature sets, 10 different models were trained and their final performance values (correlations) were calculated. Then, I took the average of the correlations.

- Prediction ensembling - Average predictions: From the feature sets, 10 different models were trained, and predictions were made for each sample. Then, I took the average predictions by sample.

- Feature ensembling: From the feature sets, each feature was concatenated by sample. Then, a model was trained on them. The drawback of this is that the feature space was 10 times larger. Unfortunately, it has a significant drawback in terms of computing time and memory usage. To mitigate this effect, a Principal Component Analysis (PCA) was performed on the concatenated feature data.

The first result was about the training of 10 different models. The gap between the performance of the best model and the worst model was up to .033. It demonstrated that the BoAW algorithm should be more robust, as making only one feature extraction may yield an unbiased and potentially misleading result. This problem of stochastic behaviour can be handled with prediction ensembling or feature ensembling. On the other hand, competitive results can be obtained with feature ensembling plus PCA. However, before applying PCA, we must ensure that the original number of features is lower than the number of data samples.

3 HMM/DNN as a Hybrid Feature Extraction Method

The following research aims to investigate methods beyond traditional machine learning. It explores hybrid methodologies that combine both traditional and modern approaches. Although the previous findings demonstrate the strengths and practical guidelines of traditional techniques, the field of computational paralinguistics could greatly benefit from investigating how hybrid models might perform [21]. A hybrid solution presents challenges in architecture design, computational management, and hyperparameter tuning. Hidden Markov Models were initially designed for frame-level classification. Deep Neural Network (DNN)s usually require a large amount of data. Computational paralinguistics requires utterance-level classification/regression and typically has a small corpora. This drawback makes hybrid methods more challenging to apply in paralinguistics [16, 21].

In Chapter 4. of the PhD thesis, a comprehensive study of the HMM/DNN hybrid model is presented. The model has two parts. The first part is the Deep Neural Network, which excels in feature extraction and non-linear mapping. The second part is a Hidden Markov Model. It handles temporal modelling. Figure 2 shows the complete flow of training and using a HMM/DNN model. First, we need to train the model on a larger ASR corpus. Once the training of the hybrid model is complete, we need to make a slight modification to our model. We have to detach the DNN from the hybrid model and fix its weights. Then, we have to focus on the output of the last few hidden layers, as they can provide more abstract information. The modification utilises the network for embedding extraction.

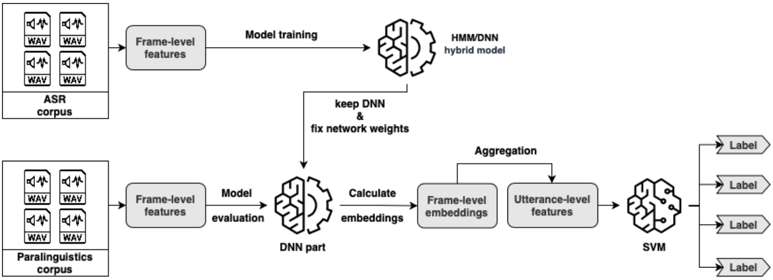

Figure 2: *Hybrid HMM/DNN Model Workflow for Paralinguistic Tasks.*

3.1 Topic independence

This research demonstrates best practices and global guidelines for combining traditional methods with deep neural networks in the case of a HMM/DNN model. On the one hand, the main focus was on exploring different feature aggregation strategies. Parallel to that, another investigation was conducted regarding the architectural choices of the DNN.

Three different databases were used to cover multiple paralinguistic use-cases: AIBO [26], URTIC [20], and iHEARu-EAT [18]. The classifier was a Support Vector Machine. For frame-level embedding extraction, we used the activation values of the five middle hidden layers (i.e., layers 1, 2, 3, 4, and 5). For the utterance-level feature vectors, the aggregation approaches used were the following: arithmetic mean, standard deviation, kurtosis, skewness, and zero ratio.

In an overall conclusion, clear guidelines can be provided on parameter configurations to effectively use the HMM/DNN hybrid technique:

- Extracting embeddings from the 4th layer always gives the best performance scores.

- Combining at least three aggregation techniques will always improve the results in any paralinguistic task. But the best combination setup may be task-dependent.
- When choosing the number of aggregations to combine, take into account Occam's razor principle.
- Always consider including the mean, standard deviation and zero ratio in the combination. Mean and standard deviation consistently performed best, while non-traditional techniques can enhance their performance.

4 Deep Neural Networks as State-of-the-Art Feature Extraction Models

In the field of computational paralinguistics, the most effective solutions nowadays integrate DNNs. There is a growing interest in general feature extractors that are non-specific to any paralinguistic tasks, such as Seq. Autoencoders and Wav2Vec 2.0s. These methodologies were initially developed for specific purposes but were later employed as frame-level feature extractors in computational paralinguistics [15, 17]. The small size of paralinguistic datasets makes it difficult to train these feature extractor models from scratch. To address this issue, usually a standard ASR corpus is used for training purposes. These models can automatically learn complex representations directly from raw audio, capturing important patterns and contextual relationships in speech data. These embeddings may outperform hand-crafted features. [25, 28].

In Chapter 5. of the PhD thesis, the focus is on two DNN: the Seq. Autoencoder model and the Wav2Vec 2.0 model. In the first research, the primary focus is on the importance of the audio preprocessing and the aggregation strategy in the case of the Seq. Autoencoder model. The second study demonstrates the use of the Wav2Vec 2.0 model as a frame-level feature extractor and explores the effects of various aggregation strategies. Guidelines are established for proper data preprocessing and aggregation techniques across diverse paralinguistic tasks.

4.1 Audio Preprocessing and Aggregation Methods

Seq. Autoencoder technique was developed for creating compressed and representative embeddings for data samples. In computational paralinguistics, these representations can be used as frame-level features.

In this study, a Hungarian Mild Cognitive Impairment database was used [5]. The frame level embeddings were extracted from a Seq. Autoencoder model. Figure 3 shows the structure of the network. The basic idea is to train a neural network to reconstruct the input, while the network structure contains a small-sized *bottleneck layer*. Using the activation values of the bottleneck layer leads to a compressed representation of the input. Here, the network was trained on a subset

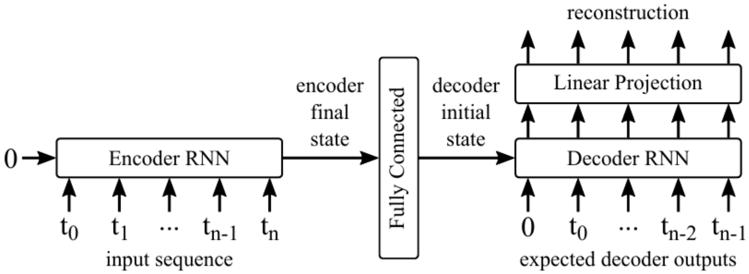

Figure 3: *An overview of the recurrent Seq. Autoencoder [12].*

of the BEA corpus [22]. After the initial training, the decoder should be detached. Only the encoder is used to map frame-level raw audio features to fixed-size embeddings. Instead of aggregating the features, the classification was performed at the level of 5-second chunks. The classifier was an SVM. The frame-level predictions were aggregated to the utterance level.

The first experiment focused on the influence of the audio preprocessing. It involved removing background noise by clipping power levels to a given dB value (-30, -45, -60, -75). The second experiment focused on different feature aggregation strategies. Four aggregation methods were evaluated on chunk-level posteriors to obtain speaker-level scores. The aggregations were: arithmetic mean, median, geometric mean, and harmonic mean.

The results highlighted practical recommendations for implementing Seq. Autoencoder feature extraction in computational paralinguistics:

- Removing background noise can increase the classification performance by clipping power levels below -75 dB.
- Among various power-level clipping strategies, individual threshold conditions outperform combined feature sets.
- The median aggregation performed better for two-class tasks, while the harmonic mean improved three-class distinctions. It is highlighting the need for task-specific aggregation selection.

These results provide valuable guidelines for both academic research and practical implementations.

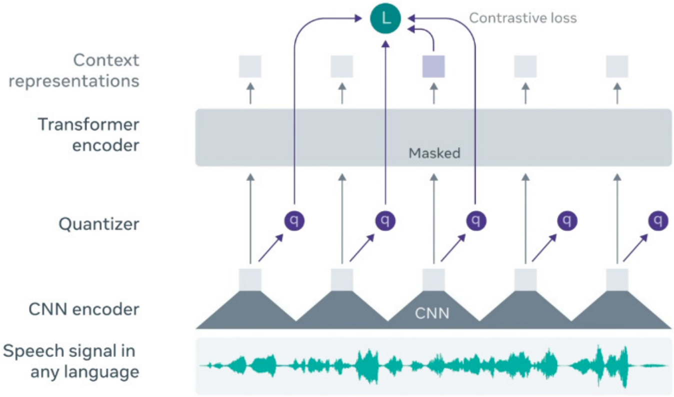

Figure 4: *The structure of the Wav2Vec 2.0 model [13].*

4.2 Aggregation Strategies

Wav2Vec 2.0 technique was developed to learn general speech feature representations by pre-training on large amounts of unlabelled speech data. This experiment has two main focuses: architectural choices and aggregation techniques.

In this experiment, 3 databases were used: AIBO [26], URTIC [20] and iHEARu-EAT [18]. To extract frame-level embeddings, a fine-tuned Wav-to-Vec 2.0 Neural Network model was used. Figure 4 shows the structure of the model. First, it has a CNN to encode the waveform into latent feature vectors. Then, it has a transformer network to generate contextualised embeddings. To investigate architectural design options, frame-level embeddings were extracted from both the final layer of the CNN and the last layer of the transformer. For utterance-level features, 11 aggregation methodologies were employed. This study challenges the conventional reliance on only the most common statistical aggregations and examines both individual and combined configurations. The classification was performed with an SVM model.

The results highlight that proper preprocessing and aggregation strategies are crucial and provide guidelines for selecting appropriate design choices:

- Convolutional embeddings behave more robustly and have global behavioural patterns in different databases.
- The mean aggregation and the middle percentile aggregations are also competitive techniques.
- The traditional standard deviation and median aggregations are heavily topic dependent.
- Wav2Vec 2.0 embeddings can contain outlier values. As a result, the minimum and maximum aggregation methods perform less robustly.
- The peak of performance fell on the combination of 3-4 aggregation techniques. The best combinations typically include the mean and a non-traditional percentile value. It can improve the generalisation ability of the model, while keeping the feature space below 2048.

These guidelines serve as valuable references for both academic research and practical implementations.

5 Summary

This PhD thesis presents comprehensive research in the field of computational paralinguistics, especially for the three main categories of machine learning approaches (traditional, deep learning-based, and hybrid methodologies). Despite the growing number of studies in this area, there is still no consensus on a set of architectural design patterns that can be applied universally. Some approaches may work well for specific datasets, yet fail to generalise across multiple use-cases.

This gap in the literature motivated my study. This thesis established global guidelines in speech-based classification and regression tasks. The following fundamental challenges were encountered through a systematic investigation of feature extraction methodologies. First of all, most paralinguistic corpora remain small (less than 100 hours), making it more challenging to observe and draw conclusions about global trends. The extremely low amount of data is also limiting the training of Deep Neural Networks. Moreover, cross-cultural generalisation is also a huge challenge. For example, models trained on Western speech underperform on tonal languages. Lastly, computational costs play a crucial role in real-life applications. Deep Neural Networks require more resources than traditional methods, making low-resource deployment more difficult. Comprehensive research in this field is crucial for the everyday development of paralinguistic systems. This thesis enhances the understanding of features within paralinguistic analysis and identifies methods that could improve the overall effectiveness of computational models.

Table 1 summarises the relation between the thesis points and the corresponding publications.

	[1]	[8]	[3]	[9]	[4]	[2]	[5]	[7]	[6]
I/1.	●	●							
I/2.			●	●					
I/3.					●				
II/1.						●			
III/1.							●	●	
III/2.									●

Table 1: *Thesis points and the corresponding publications.*

Contributions of the thesis

In the **first thesis group**, my contributions are related to end-to-end experiment pipelines and result analysis. Detailed discussion can be found in Chapter 3.

I / 1. I had implemented the end-to-end pipeline from data preprocessing to classification, while taking care of the experimental setups. I also performed a statistical comparison of configurations and identified optimal parameter settings while documenting the findings.

I / 2. I had implemented an experimental framework for analysing corpus independence in BoAW feature extraction. I had preprocessed databases, constructed cross-corpus codebooks, and conducted systematic tests, while documenting all the results and conclusions.

I / 3. I had implemented the experimental framework for the stochastic variability, while documenting all of the results. I had developed an infrastructure for testing feature extractions with multiple random seeds. I had implemented various aggregation strategies and addressed the data dimensionality issue using PCA.

In the **second thesis group**, my contributions are related to aggregation pipelines and result analysis. Detailed discussion can be found in Chapter 4.

II / 1. I had implemented an aggregation pipeline, with various aggregations, feature transformations, and classification. I had conducted multiple iterations of experiments to analyse the effect of layers and aggregations. I had identified optimal parameter settings while also documenting the findings.

In the **third thesis group**, my contributions are related to data preprocessing, aggregation and classification pipelines. Detailed discussion can be found in Chapter 5.

III / 1. I had implemented a classification pipeline that included various prediction-level aggregations. Based on the already calculated posterior values, I have conducted multiple iterations of experiments to analyse the effect of noise reduction and aggregation. I had performed a comparison of configurations and documented my findings.

III / 2. I had implemented a classification pipeline. Based on already calculated frame-level feature vectors, I performed data preparation and integrated the output into an SVM-based classification model. I had identified optimal layer settings and aggregation techniques and documented my findings.

The author's publications on the subjects of the thesis

Journal publications

[1] Vetráb, M. & Gosztolya, G. (2022). Using the Bag-of-Audio-Words approach for emotion recognition. Acta Universitatis Sapientiae, Informatica, 14(1), 2022. 1-21. https://doi.org/10.2478/ausi-2022-0001

[2] Vetráb, Mercedes & Gosztolya, Gábor. (2023). Using Hybrid HMM/DNN Embedding Extractor Models in Computational Paralinguistic Tasks. Sensors. 23. 5208; - https://doi.org/10.3390/s23115208

Full papers in conference proceedings

[3] Vetráb, M., Gosztolya, G. (2020). Investigating the Corpus Independence of the Bag-of-Audio-Words Approach. In: Text, Speech, and Dialogue. TSD 2020. Lecture Notes in Computer Science(), vol 12284. Springer

https://doi.org/10.1007/978-3-030-58323-1_31

[4] M. Vetráb and G. Gosztolya, Handling the stochastic behaviour of the Bag-of-Audio-Words method, 2022 IEEE 20th Jubilee World Symposium on Applied Machine Intelligence and Informatics (SAMI), Slovakia, 2022, pp. 000021-000026

https://doi.org/10.1109/SAMI54271.2022.9780776

[5] M. Vetráb et al., Using Spectral Sequence-to-Sequence Autoencoders to Assess Mild Cognitive Impairment, In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 6467-6471

https://doi.org/10.1109/ICASSP43922.2022.9746148

[6] Vetráb, M., Gosztolya, G. (2023). Aggregation Strategies of Wav2vec 2.0 Embeddings for Computational Paralinguistic Tasks. In: Speech and Computer. SPECOM 2023. Lecture Notes in Computer Science(), vol 14338. Springer

https://doi.org/10.1007/978-3-031-48309-7_7

[7] Kiss-Vetráb, M. és José Vicente, E. és Balogh, R. és Imre, N. és Hoffmann, I. és Tóth, L. és Pákáski, M. és Kálmán, J. és Gosztolya, G. (2022) Enyhe kognitív zavar automatikus felismerése szekvenciális autoenkóder használatával. In: XVIII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, pp. 175-184. ISBN 9789633068489

[8] Mercedes, V., & Gábor, G. (2019). Érzelmek felismerése magyar nyelvű hangfelvételekből akusztikus szózsák jellemzőreprezentáció alkalmazásával. In: XV. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem Informatikai Intézet, Szeged, pp. 265-274. ISBN: 9789633153932

[9] Vetráb, Mercedes és Gosztolya, Gábor (2020) Az akusztikus szózsák eljárás korpuszfüggetlenségének vizsgálata. In: XVI. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, pp. 219-231. ISBN 9789633067192

Further related publications

[10] Gosztolya, G., Vetráb, M., Svindt, V., Bóna, J., & Hoffmann, I. (2024). Wav2vec 2.0 Embeddings Are No Swiss Army Knife – A Case Study for Multiple Sclerosis. Interspeech 2024, 2499–2503.

https://doi.org/10.21437/Interspeech.2024-995

[11] Egas-López, J. V., Vetráb, M., Tóth, L., & Gosztolya, G. (2021b). Identifying Conflict Escalation and Primates by Using Ensemble X-Vectors and Fisher Vector Features. Interspeech 2021, 476–480.

https://doi.org/10.21437/Interspeech.2021-1173

Other References

[12] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller. Sequence-to-sequence autoencoders for unsupervised representation learning from audio. In *Proceedings of DCASE*, pages 17–21, 2017.

[13] Alexei Baevski, Michael Auli, and Alexis Conneau. Wav2vec 2.0: Learning the structure of speech from raw audio, 09 2020.

[14] Felix Burkhardt, Astrid Paeschke, M. Rolfes, Walter Sendlmeier, and Benjamin Weiss. A database of German emotional speech. In *Proceedings of Interspeech*, pages 1517–1520, 09 2005.

[15] José Vicente Egas-López and Gábor Gosztolya. Deep Neural Network embeddings for the estimation of the degree of sleepiness. In *IEEE International Conference on*

Acoustics, Speech and Signal Processing, ICASSP, pages 7288–7292, 06 2021.

[16] Gábor Gosztolya, András Beke, and Tilda Neuberger. Differentiating laughter types via HMM/DNN and probabilistic sampling. In *Speech and Computer, SPECOM 2019*, volume 11658, pages 122–132, 07 2019.

[17] Gábor Gosztolya, Tamás Grósz, Róbert Busa-Fekete, and László Tóth. Detecting the intensity of cognitive and physical load using adaboost and deep rectifier neural networks. In *Proc. Interspeech 2014*, pages 452–456, 09 2014.

[18] Simone Hantke, Felix Weninger, Richard Kurle, Fabien Ringeval, Anton Batliner, Amr Mousa, and Björn Schuller. I hear you eat and speak: Automatic recognition of eating condition and food type, use-cases, and impact on asr performance. *PLOS ONE*, 11:1–24, 05 2016.

[19] Florian Hönig, Anton Batliner, Elmar Nöth, Sebastian Schnieder, and Jarek Krajewski. Acoustic-prosodic characteristics of sleepy speech - between performance and interpretation. In *Speech Prosody 2014*, pages 864–868, 2014.

[20] Jarek Krajewski, Sebastian Schieder, and Anton Batliner. Description of the upper respiratory tract infection corpus (urtic). In *Proc. Interspeech 2017*, 01 2017.

[21] N. Morgan and H. Bourlard. Continuous speech recognition using multilayer perceptrons with hidden markov models. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, volume 1, pages 413–416, 04 1990.

[22] Tilda Neuberger, Dorottya Gyarmathy, Tekla Etelka Gráczi, Viktória Horváth, Mária Gósy, and András" Beke. Development of a large spontaneous speech database of agglutinative Hungarian language. In *Proceedings of Text, Speech and Dialogue*, volume 8655, pages 424–431, 09 2014.

[23] Björn Schuller and Anton Batliner. *Computational paralinguistics: emotion, affect and personality in speech and language processing*. Wiley Publishing, 11 2013.

[24] Björn Schuller, Stefan Steidl, and Anton Batliner. The INTERSPEECH 2009 emotion challenge. In *Proc. Interspeech 2009*, pages 312–315, 01 2009.

[25] Björn W. Schuller, Anton Batliner, Christian Bergler, and at all. The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates. In *Proc. Interspeech 2021*, pages 431–435, 01 2021.

[26] Stefan Steidl. *Automatic classification of emotion related user states in spontaneous children's speech*. Logos-Verlag Berlin, Germany, 05 2009.

[27] Dávid Sztahó, Viktor Imre, and Klára Vicsi. Automatic classification of emotions in spontaneous speech. In *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues*, volume 6800, pages 229–239, 09 2011.

[28] Johannes Wagner, Dominik Schiller, Andreas Seiderer, and Elisabeth Andre. Deep Learning in Paralinguistic Recognition Tasks: Are Hand-crafted Features Still Relevant? In *Proc. Interspeech 2018*, pages 147–151, 09 2018.

6 Összefoglalás

A doktori értekezés célja, hogy átfogó kutatást nyújtson a számítógépes paralingvisztika területén, magába foglalva a gépi tanulás három fő kategóriáját (tradicionális, mélytanuláson alapuló és hibrid módszerek). Annak ellenére, hogy egyre több tanulmány készül a területen, még mindig nincs konszenzus az általánosan alkalmazható architektúrális tervezési mintákról. Egyes megközelítések sokszor csak bizonyos adatbázisok esetében működnek jól. Ez, a szakirodalomban fennálló hiányosság motiválta kutatásomat. A disszertáció célja, hogy olyan globális irányelveket határozzon meg melyek diverz feladatok esetében is optimális megoldásokhoz vezetnek.

A disszertáció 4 fő témakörből áll. Az első fejezet bemutatja a legfontosabb alapvető fogalmakat, beleértve a technikai kihívásokat, a használt adatbázisokat és az alkalmazott módszereket. A második fejezet a BoAW tradicionális módszerrel elért eredményeket mutatja be. Részletezi, a teljesítményét különböző paralingvisztikai feladatokon, és a paraméteroptimalizálás fontosságát. A harmadik fejezet bemutatja a hibrid HMM/DNN rendszert, amely ötvözi a tradicionális statisztikai és a modern mélytanulási megközelítéseket. Tárgyalja, a megfelelő hangelőfeldolgozás és a jó aggregációs stratégia fontosságát A negyedik fejezet két mélytanuló modellt mutat be: a Szekvenciális Autoencodert és a Wav2Vec 2.0 modellt. A kísérletek fókusza az audio előfeldolgozás optimalizálásán és fejlett aggregációs stratégiákon van.

Minden kísérlet fő aspektusa, a robusztus, általánosítható megoldásokra való törekvés. Jelentőségük abban rejlik, hogy paralingvisztikai alkalmazások tekintetében, gyakorlati iránymutatásokat nyújtanak a különböző gépi tanulási technikák implementálásához.

DECLARATION

In the PhD dissertation of Mercedes Kiss-Vetráb entitled „Computational Paralinguistics: The Importance of Audio Analysis and Feature Extraction Methodologies", the contribution of Mercedes Kiss-Vetráb was decisive in the following results:

Thesis point I/1

Relevant publication: Mercedes Vetráb and Gábor Gosztolya. Aggregation strategies of wav2vec 2.0 embeddings for computational paralinguistic tasks. In International Conference on Speech and Computer, pages 79–93. Springer Nature Switzerland Cham, 2023.

Thesis point I/1

Relevant publication: Mercedes Vetráb and Gábor Gosztolya. Érzelmek felismerése magyar nyelvű hangfelvételekből akusztikus szózsák jellemzőreprezentáció alkalmazásával. XV. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem Informatikai Intézet, Szeged, pages 265–274, 2019

Thesis point I/2

Relevant publication: Mercedes Vetráb and Gábor Gosztolya. Az akusztikus szózsák eljárás korpuszfüggetlenségének vizsgálata. 2020. In XVI. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, pp. 219-231. ISBN 9789633067192

Thesis point I/2

Relevant publication: Mercedes Vetráb and Gábor Gosztolya. Investigating the corpus independence of the bag-of-audio-words approach. In Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23, pages 285–293. Springer International Publishing, 2020.

Thesis point I/3

Relevant publication: Mercedes Vetráb and Gábor Gosztolya. Handling the stochastic behaviour of the bag-of-audio-words method. In 2022 IEEE 20th Jubilee World Symposium on Applied Machine Intelligence and Informatics (SAMI), pages 000021–000026.

Thesis point II/1

Relevant publication: Mercedes Vetráb and Gábor Gosztolya. Using hybrid HMM/DNN embedding extractor models in computational paralinguistic tasks. In Sensors, 23(11):5208, 2023.

Thesis point III/1

Relevant publication: Mercedes Vetráb, José Vicente Egas-López, Réka Balogh, Nóra Imre, Ildikó Hoffmann, László Tóth, Magdolna Pákáski, János Kálmán, and Gábor Gosztolya. Using spectral sequence-to-sequence autoencoders to assess mild cognitive impairment. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6467–6471. IEEE, 2022.

Thesis point III/1

Relevant publication: Mercedes Vetráb, José Vicente Egas-López, Réka Balogh, Nóra Imre, Ildikó Hoffmann, László Tóth, Magdolna Pákáski, János Kálmán, and Gábor Gosztolya. Enyhe kognitív zavar automatikus felismerése szekvenciális autoenkóder használatával. 2022. In XVIII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, pp. 175-184. ISBN 9789633068489

Thesis point III/2

Relevant publication: Mercedes Vetráb and Gábor Gosztolya. Aggregation strategies of wav2vec 2.0 embeddings for computational paralinguistic tasks. In International Conference on Speech and Computer, pages 79–93. Springer Nature Switzerland Cham, 2023.

These results cannot be used to obtain an academic research degree, other than the submitted PhD thesis of Mercedes Kiss-Vetráb.

Date: 2025, August 25

.... (signature of candidate)

.... (signature of supervisor)

The head of the Doctoral School of Computer Science declares that the declaration above was sent to all of the coauthors and none of them raised any objections against it.

Date: 2025, August

SZEGEDI TUDOMÁNYEGYETEM
INFORMATIKA DOKTORI ISKOLA
SZEGED

....

(signature of head of Doctoral School)