S M Asif Hossain

Student ID: Y935M825

Professor Huabo Lu

CS 800

Homework 4

March 07 2025

Speed-Based Stress Classification Based on Modulation Spectral Features and Convolutional Neural Networks

In this paper, Kshirsagar et al. studied how stress can be detected from human speech [1]. Stress is harmful as it can lead to serious health issues like heart problems, depression, or sleeping troubles. To detect stress from speech, the authors used something called modulation spectral features (MSF). These features measure changes in how a person's voice sounds when they are stressed. First, they adjusted the loudness of speech samples. Next, they divided the speech into different frequency channels similar to how human ears hear sound. They then calculated how speech patterns changed over time. A mathematical tool called the Hilbert transform was used to get the temporal envelope of speech, which helps measure these changes.

For classifying stress levels from these features, the authors used a Convolutional Neural Network (CNN). They compared their CNN-based method to simpler methods like Support Vector Machines (SVM) and another type of neural network called Deep Neural Network (DNN). These other methods used standard speech features extracted by OpenSMILE.

The study used a database called SUSAS, which includes speech samples under nine stress conditions like neutral, angry, loud, soft, fast, slow, Lombard (speech under noise), and high or low workload conditions. In the experiments, the CNN method showed the highest accuracy, which is 70% in classifying speech into these nine stress categories. This was significantly better than the simpler methods, which achieved lower accuracy.

The researchers found that their CNN approach worked better as the complexity of classification increased. They discussed that CNN was successful because it automatically learns the most relevant features without depending too much on manual feature selection. However, when tasks had fewer categories like only neutral versus angry speech, the simpler methods performed similarly or even better than the CNN.

Modulation Spectral Signal Representation and I-Vectors For Anomalous Sound Detection

In this paper, Kshirsagar et al. conducted research on detecting unusual sounds in machines using two main methods [5]. Detecting these sounds can help identify machine problems early, preventing costly breakdowns and production issues. The first method used a special feature called modulation spectral features (MSF), which captures changes in sound patterns over time. These features are similar to how human ears process sounds. The authors processed the sound signals to remove noise, adjusted loudness, and divided the sounds into frequency channels. A mathematical approach called the Hilbert transform was then used to find patterns in these sounds. The formula used to calculate the temporal envelope from the Hilbert transform is:

$$e_j(n) = \sqrt{\hat{x}_j(n)^2 + H\{\hat{x}_j(n)\}^2}$$

Here, $e_j(n)$ is the temporal envelope of the sound signal, $x_j(n)$ is the processed speech signal, and $H\{x_j(n)\}$ is the Hilbert transform applied to the signal.

In the first method, authors grouped normal sounds using graph-based clustering. This involved creating clusters of normal machine sounds and calculating how far a new sound was from these clusters to identify anomalies. Sounds far from these normal clusters were marked as unusual.

The second method used mel-frequency cepstral coefficients (MFCC) combined with a model called i-vectors. I-vectors summarize sound data into fixed-size vectors, making it easier to analyze. These vectors were then fed into a Gaussian Mixture Model (GMM), which calculates the likelihood of a sound being normal. Sounds with low likelihoods were considered anomalous. The equation for extracting mel-frequency cepstral coefficients is:

$$c_n = \sum_{m=1}^{M} [Y_m] \cos\left[\frac{\pi n}{M}\left(m - \frac{1}{2}\right)\right], \quad n = 1, 2, 3, ..., N$$

Here, c_n is the cepstral coefficient, Y_m represents the log-energy of the m-th frequency band, M is the total number of frequency bands, and N is the total number of coefficients.

The authors tested their methods on different machines, comparing their results to standard methods. They found their approaches significantly improved anomaly detection, particularly when they combined both methods. Their combined system improved detection accuracy by about 11% compared to traditional methods. The authors concluded that using MSF and i-vector-based methods together effectively identifies anomalous machine sounds.

Modulation Spectral Signal Representation for Quality Measurement and Enhancement of Wearable Device Data: A Technical Note

In this paper, Kshirsagar et al. introduced a new method called modulation spectral signal representation to improve data quality collected from wearable devices [9]. Wearable devices like smartwatches and health trackers provide important health data such as heart rate and movement. However, these devices often capture noisy data due to movements or environmental disturbances. This noise reduces the accuracy and usefulness of the collected data, making it harder to reliably monitor health or detect diseases.

The authors tried to solve this problem by using modulation spectral features (MSF). MSF helps identify useful signals by measuring how quickly the frequency content of the data changes over time. This method works similarly to human ears naturally separate sounds, allowing better separation between actual signals and noise. The team processed the wearable device data to reduce noise, normalize loudness, and break it into frequency bands.

The authors tested their method on different wearable signals, including heart rate, ECG, speech, and EEG signals. Their results showed that MSF significantly improved the quality of data. For ECG signals, their method was effective in clearly separating heartbeat signals from noise, improving the accuracy of heart rate monitoring. Similarly, it improved speech signal clarity and reduced noise effectively in EEG data. The authors concluded that the MSF method greatly improves the reliability of wearable device data, making it highly useful for healthcare and clinical research.

COVID-19 Detection via Fusion of Modulation Spectrum and Linear Prediction Speech Features

In this paper, Kshirsagar et al. conducted research on using speech signals to detect COVID-19 infections [10]. Detecting COVID-19 quickly and easily is important, but common testing methods are expensive, slow, or less reliable. The authors focused on speech analysis because COVID-19 affects the lungs, throat, and vocal cords, causing noticeable changes in a person's voice.

The researchers introduced two main methods to analyze speech signals. The first method uses modulation spectral features (MSF), which look at how speech frequency changes over time. This helps in identifying unique voice patterns caused by COVID-19. They processed the speech signals by converting them into spectrograms, which visually show the changes in frequency over time. They then transformed these into modulation spectrograms, clearly highlighting differences between normal and COVID-19 speech.

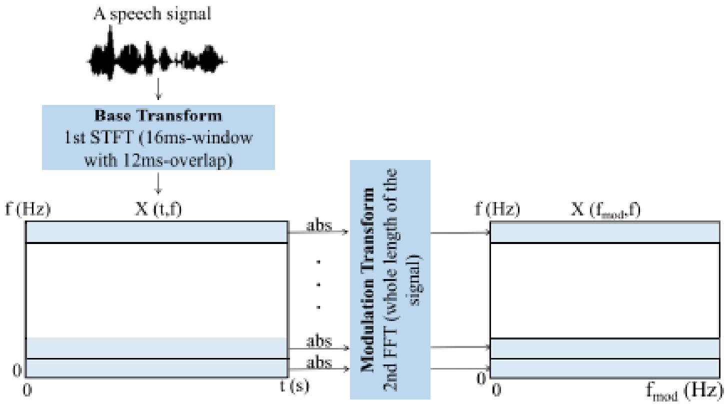

Figure 1: Signal processing steps involved in the computation of the modulation spectrogram

The second method is called Linear Prediction (LP) analysis, which separates speech into two components: the vocal tract signal and the excitation signal. The vocal tract signal shows how the shape of the mouth and throat changes during speech. The excitation signal relates to how sound is produced by the vocal cords.

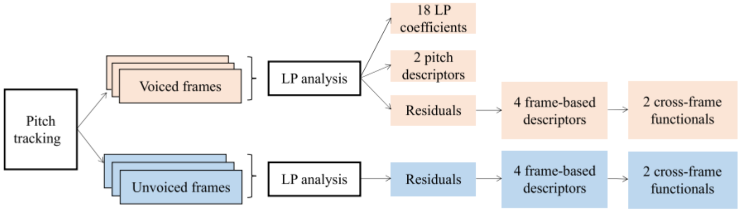

Figure 2: Signal processing steps involved in the computation of the vocal tract and excitation signals from LP analysis

The authors developed a two-stage classification system. In the first stage, speech samples are checked using MSF features. Samples initially identified as negative are further checked using LP features in the second stage. This two-step process significantly improved detection accuracy.

They tested their system using three different speech datasets. The results showed their method worked better than other existing systems, achieving higher accuracy and reliability.

Towards Robust Building Damage Detection: Leveraging Augmentation and Domain Adaptation

In this paper, Kshirsagar et al. conducted research on detecting building damage using satellite images after natural disasters [4]. Detecting damage quickly and accurately helps authorities respond faster and better. Current methods often struggle to clearly identify buildings' edges and corners, making it difficult to classify damage accurately.

The authors introduced a new method using data augmentation and domain adaptation to improve building damage detection. They used a fusion-based augmentation technique that clearly highlights important features such as building edges and corners. This method combines edge detection, contrast enhancement, and unsharp masking. Edge detection identifies sharp changes in the image, helping highlight structural details. Contrast enhancement makes objects stand out clearly against the background and unsharp masking improves the overall sharpness of the images.

To address differences between datasets, they used domain adaptation techniques. Domain adaptation helps models trained on one set of images work well on a different set. The authors used two methods, supervised fine-tuning and unsupervised deep CORAL. Supervised fine-tuning uses a small amount of labeled data from the target domain to adapt the model. Unsupervised deep CORAL aligns the features of both datasets without needing labeled target data.

Their experiments showed significant improvements. The proposed augmentation increased accuracy by about 5-7% for minor and major damage detection and improved building localization accuracy by around 2.5%. The authors concluded that their methods effectively enhanced the detection and classification of building damage, making disaster response quicker and more reliable.

Ad-Hoc Monitoring of COVID-19 Global Research Trends for Well-Informed Policy Making

In this paper, Sarkar et al. conducted research to monitor global COVID-19 research trends effectively to assist policymakers [7]. With the rapid spread of COVID-19, many researchers around the world have quickly published studies related to the pandemic. Due to the large number of publications, it became challenging for policymakers to keep track of relevant research effectively.

To solve this issue, the authors introduced a method called Ad-Hoc Topic Tracking. This approach helps policymakers track research based on their specific interests without needing to manually review all papers. The method has two main parts, Zero-Shot Topic Categorization and Spatio-Temporal Analysis.

Zero-Shot Topic Categorization automatically sorts research papers into user-defined topics based on keywords provided by the users. This method uses advanced language processing models to categorize articles immediately without needing prior training on specific topics. Spatio-Temporal Analysis examines how these categorized topics evolve over time and across different regions. It uses visual tools like interactive maps and graphs to clearly display where and when certain research topics become popular.

Users first input research papers and select their topics of interest. Then, the zero-shot categorization method labels each paper with relevant topics. Finally, the labeled papers are visualized using interactive maps and graphs to show how research topics evolve globally over time.

The researchers tested their system using thousands of COVID-19 research articles. Their results showed that their method accurately categorized topics and clearly visualized global research trends. Policymakers found this system helpful for quickly understanding research trends and making informed decisions during the pandemic. The authors concluded their method greatly supports policymakers by providing clear, timely, and customizable insights into global COVID-19 research.

Exploring Universal Sentence Encoders for Zero-shot Text Classification

In this paper, Sarkar et al. explored the effectiveness of Universal Sentence Encoders (USE) for zero-shot text classification [8]. Zero-shot classification is a challenging task because it involves classifying text into categories without having training examples beforehand. This approach is helpful because collecting labeled data can be expensive and time-consuming.

In their study, the authors explored two architectures of USE, Transformer-based and DAN-based (Deep Averaging Network). They compared these with a simpler topic-based classification approach called Generative Feature Language Models (GFLM). To perform zero-shot classification, the authors converted text articles and category labels into embeddings using USE. They then measured the similarity between article embeddings and label embeddings. Articles were assigned labels based on the similarity score exceeding a set threshold.

The authors tested their methods using seven datasets covering various topics such as medical articles, news, and product reviews. They found that the simpler GFLM methods often performed better than USE-based methods, especially when categories had many similar meanings. The DAN-based USE performed slightly better than Transformer-based USE, and using explicit text for label embeddings provided more accurate results than using only label names and keywords.

The authors discussed that USE struggled in cases where labels had high semantic overlap, meaning labels had very similar meanings. They discovered that reducing the number of overlapping labels improved USE performance. They concluded that while USE is popular and useful, simpler topic-based methods might perform better for zero-shot text classification when categories are highly similar.

COVID19α: Interactive Spatio-Temporal Visualization of COVID-19 Symptoms through Tweet Analysis

In this paper, Sarker et al. conducted research on visualizing COVID-19 symptoms using data from Twitter [2]. The main goal was to help health experts and policymakers understand how COVID-19 symptoms were spreading globally through people's tweets. Social media, especially Twitter, became an important source of real-time information during the pandemic.

The authors developed an interactive visualization tool called COVID19α. This tool helps analyze and visualize COVID-19 symptoms reported by users worldwide through tweets. They collected and analyzed about 462 million tweets from March 19, 2020, to September 15, 2020. The collected tweets included user-generated descriptions of COVID-19 symptoms.

COVID19α provides three types of visualizations. The first is spatial visualization, showing differences in symptom reports between countries. For example, users can select two countries and see a word cloud comparison of the most common symptoms tweeted in those locations. Another spatial visualization method uses maps to show clusters of tweets mentioning symptoms. These maps dynamically group tweets as users zoom in or out, clearly showing symptom distribution in different places.

The second type of visualization is temporal, showing how symptom mentions change over time in a specific location. Users can interactively explore trends of various symptoms during different periods. The third type, spatio-temporal visualization, combines both time and location. Users can observe how specific symptoms changed in popularity across different regions and times.

The authors concluded that their interactive visualization tool effectively helps experts quickly understand and monitor symptom trends globally, improving their ability to respond to public health crises.

SimPal: Towards a Meta-Conversational Framework to Understand Teacher's Instructional Goals for K-12 Physics

In this paper, Sarkar et al. conducted research aimed at helping K-12 physics teachers better customize simulations for their classes [3]. Often simulations come with built-in artificial intelligence (AI) agents designed to support students during experiments. However, many teachers find it challenging to adjust these AI agents because it usually requires advanced technical knowledge they do not have.

To address this problem, the authors developed SimPal, a conversational tool powered by large language models (LLMs) such as ChatGPT-3.5 and PaLM 2. SimPal allows teachers to naturally communicate their instructional goals without technical expertise. Teachers simply describe their teaching goals in a conversation. SimPal then identifies relevant physics variables and relationships that match these goals. These variables are then turned into symbolic representations, making it easier for the original AI agents to provide targeted student support.

In their research, the authors tested SimPal on 63 physics simulations from popular platforms, including PhET and Golabz. They carefully evaluated how accurately SimPal identified the correct physics variables based on teachers' instructional goals. The researchers experimented with different types of prompts to guide SimPal, finding that clearly organized instructions provided better results. They compared the performance of two different large language models, ChatGPT-3.5 and PaLM 2, and discovered that ChatGPT-3.5 generally achieved higher accuracy.

The authors discussed how their framework helps teachers create customized teaching materials easily. They concluded that SimPal effectively helps teachers clearly define and communicate their instructional goals, making simulation-based teaching more accessible and effective without needing extensive technical skills.

Processing Natural Language on Embedded Devices: How Well Do Transformer Models Perform?

In this paper, Sarkar et al. conducted research to explore how well transformer-based language models perform on embedded devices [6]. Transformers, such as BERT and its smaller variants like DistilBERT and TinyBERT, are commonly used in natural language processing (NLP) tasks because of their high accuracy. These models are usually large and require significant computational resources, which makes it difficult to run them efficiently on devices with limited hardware resources, like smartphones or small computers.

The researchers evaluated these transformer models on different embedded devices, including Raspberry Pi, Jetson Nano, UP2, and UDOO Bolt. They used tasks like Intent Classification (IC), Sentiment Classification (SC), and Named Entity Recognition (NER) to test performance. The goal was to see how well these models could perform complex language tasks while still meeting the constraints of limited memory and processing power.

In their study, the authors also experimented with reducing model sizes through pruning techniques. Pruning means removing some parts of the model that are less important to decrease its size and computational load. They found that reducing model layers and pruning attention heads significantly lowered model size but caused accuracy losses, particularly for complex tasks. For instance, reducing model size by about 60% could decrease accuracy by nearly 50%.

The results also showed that simpler tasks like Intent Classification were well-handled by smaller models. In contrast, more complex tasks like multi-label Sentiment Classification and Named Entity Recognition saw bigger drops in accuracy. Additionally, the researchers found that using GPUs improved inference speed but did not necessarily decrease energy use.

References

[1] Anderson R. Avila, Shruti R. Kshirsagar, Abhishek Tiwari, Daniel Lafond, Douglas O'Shaughnessy, and Tiago H. Falk. Speech-based stress classification based on modulation

spectral features and convolutional neural networks. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5, 2019.

[2] Biddut Sarker Bijoy, Syeda Jannatus Saba, Souvika Sarkar, Md Saiful Islam, Sheikh Rabiul Islam, Mohammad Ruhul Amin, and Shubhra Kanti Karmaker Santu. Covid19: Interactive spatio-temporal visualization of covid-19 symptoms through tweet analysis. In *Companion Proceedings of the 26th International Conference on Intelligent User Interfaces*, IUI '21 Companion, page 28–30, New York, NY, USA, 2021. Association for Computing Machinery.

[3] Effat Farhana, Souvika Sarkar, Ralph Knipper, Indrani Dey, Hari Narayanan, Sadhana Puntambekar, and Santu Karmaker. Simpal: Towards a meta-conversational framework to understand teacher's instructional goals for k-12 physics. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, L@S '24, page 461–465, New York, NY, USA, 2024. Association for Computing Machinery.

[4] Bharath Chandra Reddy Parupati, Shruti Kshirsagar, Rajiv Bagai, and Atri Dutta. Towards robust building damage detection: Leveraging augmentation and domain adaptation.

[5] William Romine, Noah Schroeder, Tanvi Banerjee, and Josephine Graft. Toward mental effort measurement using electrodermal activity features. *Sensors*, 22(19), 2022.

[6] Souvika Sarkar, Mohammad Fakhruddin Babar, Md Mahadi Hassan, Monowar Hasan, and Shubhra Kanti Karmaker Santu. Processing natural language on embedded devices: How well do transformer models perform? In *Proceedings of the 15th ACM/SPEC International Conference on Performance Engineering*, ICPE '24, page 211–222, New York, NY, USA, 2024. Association for Computing Machinery.

[7] Souvika Sarkar, Biddut Sarker Bijoy, Syeda Jannatus Saba, Dongji Feng, Yash Mahajan, Mohammad Ruhul Amin, Sheikh Rabiul Islam, and Shubhra Kanti Karmaker ("Santu").

Ad-hoc monitoring of covid-19 global research trends for well-informed policy making. *ACM Trans. Intell. Syst. Technol.*, 14(2), February 2023.

[8] Souvika Sarkar, Dongji Feng, and Shubhra Kanti Karmaker Santu. Exploring universal sentence encoders for zero-shot text classification. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang, editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 135–147, Online only, November 2022. Association for Computational Linguistics.

[9] Abhishek Tiwari, Raymundo Cassani, Shruti Kshirsagar, Diana P. Tobon, Yi Zhu, and Tiago H. Falk. Modulation spectral signal representation for quality measurement and enhancement of wearable device data: A technical note. *Sensors*, 22(12), 2022.

[10] Yi Zhu, Abhishek Tiwari, João Monteiro, Shruti Kshirsagar, and Tiago Henrique Falk. Covid-19 detection via fusion of modulation spectrum and linear prediction speech features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1536–1549, 2023.