Audio Pattern Recognition: Robust Classification of Speech vs. Urban Noise using Spectral and Temporal Descriptors

Simon Ferrier
Erasmus Student at Dipartimento di Computer Science
Università degli Studi di Milano, Milan, Italy
Home Institution: *Polytech Dijon, Université de Bourgogne*, France
Email: simon_ferrier@etu.u-bourgogne.fr

***Abstract*—This paper addresses the challenge of automatic audio classification in urban environments, specifically distinguishing human speech from background noise. The proposed methodology leverages features from multiple domains, including Mel-Frequency Cepstral Coefficients (MFCCs) for spectral texture and energy for temporal impulsivity. We evaluate an unsupervised approach using K-means clustering to visualize the feature space, followed by a comparative study of supervised classifiers, namely k-Nearest Neighbors (k-NN) and Support Vector Machines (SVM). Our findings indicate that the SVM with a Radial Basis Function (RBF) kernel achieves the highest performance with a mean accuracy of 90.50%. Furthermore, we demonstrate that the model's decisions are physically explainable through the analysis of acoustic variance and signal transients, meeting the requirements for transparent and reliable environmental monitoring systems.**

I. Introduction

AUTOMATIC speech detection in noisy environments is a fundamental task in audio pattern recognition. While traditional systems focus on general sound classification, the specific discrimination of human voice from pervasive background noise has become crucial for the development of smart wearable devices. The importance of this task lies in its ability to bridge the gap between acoustic comfort and situational awareness.

A primary potential application for this technology is the enhancement of Active Noise Cancellation (ANC) systems. Current ANC technologies are highly effective at mitigating constant mechanical sounds but often isolate the user from essential vocal communications. By integrating a robust voice detection trigger, future "Smart ANC" systems could automatically transition into transparency mode upon detecting a human voice, ensuring the user remains connected to their social environment without manual intervention.

The task presented in this study, conducted during an Erasmus mobility program at the *Università degli Studi di Milano*, is closely aligned with the objectives of the **DCASE** (Detection and Classification of Acoustic Scenes and Events) challenges. DCASE has established international benchmarks for environmental sound monitoring, specifically in Sound Event Detection (SED) and Acoustic Scene Classification (ASC). Research in this field has traditionally relied on heavy deep learning architectures; however, this project presents an alternative approach based on a multi-domain feature extraction pipeline designed for computational efficiency and real-time triggering.

The motivation behind this specific approach is to favor "explainable" features over black-box models, allowing for a better understanding of the decision boundaries in high-stress acoustic settings. We focus on two key descriptors:

1) **Spectral Domain (MFCC 2):** To capture the phonemic variance and timbre of the speech signal.
2) **Temporal Domain (Energy Ratio):** To quantify the signal's RMS energy variance, effectively separating the "bursty" nature of speech from the continuous flow of mechanical backgrounds.

The point of this study lies in the specific combination of these two low-complexity descriptors to serve as a high-speed trigger for ANC transparency. We evaluate the system's robustness by simulating speech degradation in dense urban acoustic environments. The following sections detail the data acquisition process, the mathematical foundation of the chosen descriptors, and a comparative analysis of the classification results.

II. System overview

A. Feature Selection and Initial Trials

A critical stage of the project involved identifying the most discriminative pair of descriptors. The first one we're going to use is MFCCs (the first 13 of them) because they are really common for speech detection. For the second feature, we conducted a study comparing the clustering performance of various feature sets using the K-means algorithm. In this test, we performed 100 independent runs of K-means for each pair of descriptors, limiting each run to 5 iterations. We then calculated the mean accuracy across these runs. The features we will try are :

- **Spectral Centroid:** It is defined as the center of 'gravity' of the spectrum and serves as a measure of spectral position

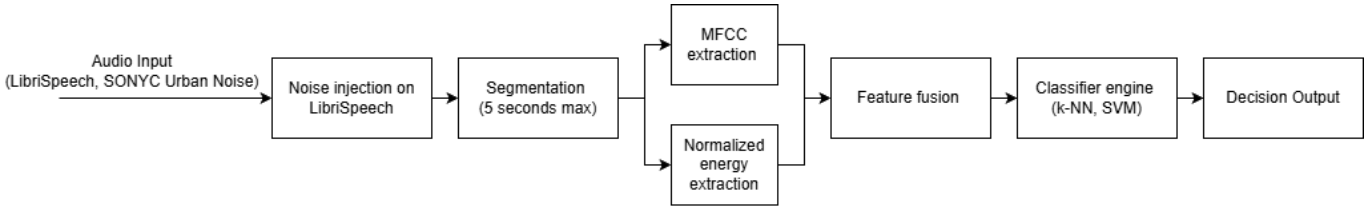

Fig. 1. Block diagram of the system

- **Spectral Rolloff:** This feature represents the frequency below which a certain percentage (usually 90%) of the magnitude distribution of the spectrum is concentrated.
- **Zero Crossing Rate:**It is the rate of sign-changes of the signal during a frame.
- **Normalized Energy:**Calculated as the ratio of RMS deviation to total RMS.

All features were normalized to ensure a zero mean and unit variance, ensuring that no single descriptor dominated the distance-based calculations. The evaluation was conducted using the **LibriSpeech** [3] dataset for voice samples and the **SONYC Urban Sound Tagging** dataset [2] for urban noise backgrounds.

The result we got from this test is the following :

TABLE I
MEAN ACCURACY OF A 100 K-MEANS WITH DIFFERENT FEATURES

Feature	Mean accuracy
Spectral Centroid	94.96%
Spectral Rolloff	92.76%
Zero Crossing Rate	95.47%
Normalized Energy	99.02%

The Energy Ratio was ultimately selected because it provided the best result in our research space. By focusing on the global temporal envelope, it remained robust against non-stationary urban noise.
The final model utilizes two complementary descriptors:

- **Spectral Domain (MFCC 2):** We extract the second Mel-Frequency Cepstral Coefficient. By calculating its standard deviation (σ), we capture the phonetic variance of speech.
- **Temporal Domain (Energy Ratio):** Defined as the coefficient of variation of the Root-Mean-Square (RMS) energy:

$$EnergyRatio = \frac{\Delta RMS}{RMS} \quad (1)$$

where RMS is the mean energy. This ratio highlights the "bursty" and rhythmic nature of human speech against the continuous energy floor of background noise.

This pair effectively balances spectral timbre information with global temporal dynamics.

B. K-mean Algorithm

Clustering by K-Mean allow us to visualize the structure of the data. It provides a geometric solution by grouping data points into K clusters. This allows us to assess whether our selected descriptors (MFCCs and Energy Ratio) create distinct clusters for "Speech" and "Urban Noise"

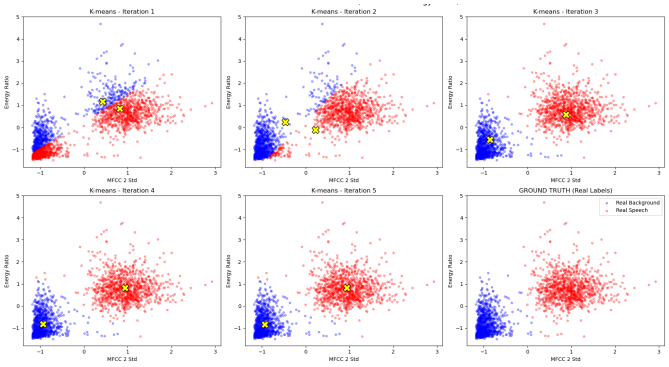

Fig. 2. K mean : Energy and MFCC

The evolution on 5 iteration show a clear separation between the two datasets. On clean data, K-means identify class with a presicion of 99% proving a good feature choice.

C. From Clean to Noisy

To avoid a biased evaluation where classification would be "too easy" under ideal conditions, we applied noise on the LibriSpeech library : This stage was crucial to ensure the system's reliability for a future Smart ANC trigger, where speech must be detected even when masked by noise.

In this experiment, a noise factor of 2 was applied during the mixing process, resulting in an induced Signal-to-Noise Ratio (SNR) of approximately -6 dB. At this level, the noise power is significantly higher than the speech signal, creating a challenging environment for detection.

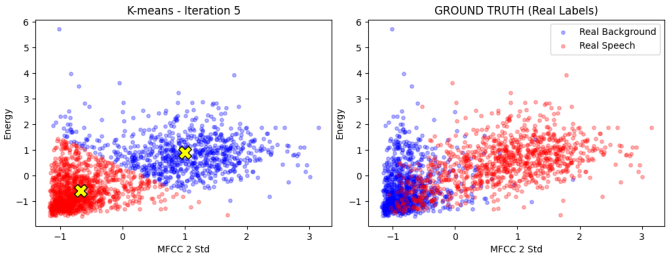

Fig. 3. Energy and MFCC with noisy signal : K-mean and ground truth comparison

With noise the K mean algorithm give a 83% mean accuracy

D. Modeling and Implementation

Two algorithmic approaches were implemented to compare their performance:

- **k-Nearest Neighbors (k-NN):** A supervised distance-based classifier that classifies a new data point by finding its 'K' closest neighbors in a labeled dataset, then assigning it the most common class (majority vote) among those neighbors, working on the principle that similar things exist near each other, using distance metrics like Euclidean distance to determine closeness.
- **Support Vector Machines (SVM):**A supervised classifier designed to find the optimal decision hyperplane that maximizes the "margin" between classes in the training data. It operates by identifying supporting hyperplanes to minimize classification errors. It can handle complex non-linear distributions by using kernel functions—such as Linear, Polynomial, or Radial Basis Function (RBF)—to map feature vectors into a higher-dimensional space where they become more easily separable.

III. Supervised Classification and Results

A. k-Nearest Neighbors (k-NN) Analysis

As outlined in the system overview, the k-NN classifier was the first supervised model evaluated. Unlike other methods, k-NN does not require a formal training stage; instead, the labeled training samples are used directly during the classification phase to find the k most similar vectors based on a distance measure. In our implementation, we used the Euclidean distance to determine the proximity between feature vectors. Since our descriptors (MFCC 2 and Energy Ratio) operate on different scales, the normalization step performed earlier was critical to ensure that the distance calculation was not biased toward a single feature.
To optimize the model, we performed a grid search across different values of k ($k \in \{1, 3, 5, 10, 20\}$).

The results of the k-NN classifier for different values of k are summarized in Table II. To ensure the robustness of our findings and account for the variability in the data splitting process, each configuration was tested over 20 independent iterations with randomized train/test sets.

TABLE II
k-NN Accuracy performance across different values of k (20-run average).

Value of k	Mean Accuracy
1	85.14%
3	90.25%
5	**90.75%**
10	89.75%
20	88.95%

The analysis of these results reveals several key insights into the model's behavior:

- **Sensitivity at low k:** For $k = 1$, the model achieves its lowest average accuracy (85.14%). This suggests a degree of overfitting, where the classifier is overly sensitive to local noise or outliers in the MFCC and Energy Ratio distributions.
- **Optimal Smoothing:** Increasing k to 5 significantly improves the performance, reaching a peak of 92.25%. By considering a larger neighborhood, the decision boundary becomes more stable, effectively filtering out minor fluctuations in the feature space.
- **Stability:** At $k = 5$, we observe not only the highest mean accuracy but also a high level of consistency across the five runs (scores ranging from 87.25% to 92.25%), indicating that the model generalizes well to unseen data.
- **Performance Saturation:** As k increases to 20, the accuracy begins to plateau (88.95%). Further increasing k would likely lead to underfitting, as the decision boundary would become too smooth, potentially ignoring the specific local characteristics of the speech and background classes.

Based on this empirical analysis, $k = 5$ was selected as the optimal parameter for our final comparison.

To understand how many sound are misunderstood by the system, we utilize the Confusion Matrix. As defined in the course, this $N_c \times N_c$ matrix provides a detailed breakdown of classification decisions by grouping results into true (ground truth) and predicted class labels. The diagonal elements represent correct classifications ($i = j$), while off-diagonal elements reveal specific types of errors. In our binary task, this tool is essential to identify if certain urban transients are being misclassified as speech, which would trigger the ANC transparency mode unnecessarily.

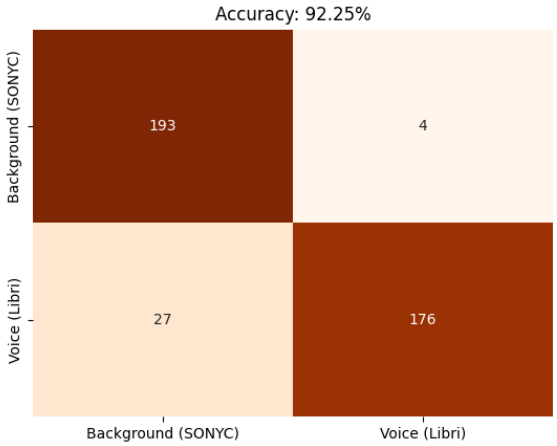

Fig. 4. k-NN confusion matrix with 5 neighbors(Best run)

When looking at which samples were misclassified, the background sound from the *SONYC* dataset that are mistaken for speech are siren and horn sounds. This confusion likely stems from their harmonic structure and sustained tonal components, which the k-NN algorithm perceives as similar to certain vocal characteristics.

Furthermore, a significant portion of *LibriSpeech* samples are classified as *Background*. This observation is consistent with our data augmentation protocol; since the speech signals were mixed with SONYC environmental noise, segments with a low LibriSpeech component volume are dominated by background features. Consequently, it is physically coherent for the classifier to prioritize the ambient noise characteristics in these specific instances.

Despite the high accuracy achieved with $k = 5$, the practical implementation of k-NN in a real-time system such as a Smart ANC wearable presents significant challenges. As highlighted in the course, the computational complexity of k-NN can be "prohibitively high" for large datasets. k-NN is a non-parametric learner that necessitates storing the entire training dataset in memory.

In the context of an embedded system, this high memory footprint and the need to calculate distances to every training point for each new audio frame could result in excessive power consumption and processing latency. This limitation makes k-NN less suitable for resource-constrained hardware compared to more compact models, despite its strong classification performance.

B. Support Vector Machines (SVM) Analysis

Following the evaluation of the k-NN, we implemented Support Vector Machines (SVM), which are considered state-of-the-art classifiers in many audio analysis tasks. Unlike the k-NN's distance-based voting, the SVM seeks to find the optimal decision hyperplane that maximizes the "margin" between our speech and urban noise classes. This approach is particularly interesting for an ANC trigger system because, once trained, the model only needs to store a small subset of the training data. The support vectors-making it potentially more efficient for real-time hardware than k-NN.

1) Choice of the kernel function: A critical part of the SVM design involves the choice of the kernel function, which maps feature vectors into a higher-dimensional "kernel space" to handle non-linear distributions. We compared four different kernels:

- **Linear Kernel:** Suitable for simple, linearly separable problems where a straight hyperplane is sufficient.
- **Polynomial Kernel:** Useful for capturing more complex relationships but requires careful tuning of the polynomial order.
- **Radial Basis Function (RBF) Kernel:** Generally the most powerful, as it can model highly non-linear decision boundaries.
- **Sigmoid Kernel:** Inspired by neural network transfer functions.

For this test we executed 20 independent SVM on each kernel function to compute the mean accuracy
We got this result :

TABLE III
SVM ACCURACY PERFORMANCE ACROSS DIFFERENT KERNEL FUNCTIONS.

Kernel Function	Mean Accuracy
Linear	89.79%
Sigmoid	80.60%
Polynomial	89.60%
RBF	**90.50%**

As shown in Table III, the high performance of the Linear kernel (89.79%) indicates that our features are highly discriminative and the classes are largely separable. However, the **RBF** kernel consistently provided the highest accuracy (90.50%). This suggests that while the classes are mostly distinct, a non-linear "curved" boundary is necessary to resolve the complex overlap where speech phonemes and urban noise share similar spectral characteristics.

2) Optimization of the Cost Parameter (C): To further refine the RBF-SVM model, we conducted a benchmark analysis of the cost parameter C. This hyperparameter is crucial as it define the penalty for misclassification: a small C prioritizes a larger decision margin, while a large C aims for a smaller training error. We evaluated $C \in \{0.1, 1, 10, 100, 1000\}$ over 20 randomized trials to ensure statistical stability. The results are detailed in Table IV.

TABLE IV
IMPACT OF THE COST PARAMETER C ON RBF-SVM ACCURACY (20-RUN AVERAGE).

C Value	Mean Accuracy
0.1	89.44%
1	90.65%
10	90.10%
100	89.27%
1000	89.12%

As shown in Table IV, the model achieves its peak performance at $C = 1$ with a mean accuracy of 90.65%. Interestingly, this configuration also provides a relatively low standard deviation (0.0113), suggesting that a wider margin helps the classifier generalize better across different noise profiles.

The fact that higher values of C (such as 1000) lead to a slight decrease in accuracy (89.12%) suggests that a stricter penalty causes the model to overfit to specific fluctuations in the urban noise background. By choosing $C = 1$, we ensure that the ANC trigger system remains robust to the non-stationary nature of environmental sounds.

IV. COMPARISON AND DISCUSSION

The comparative analysis of the implemented models highlights a significant trade-off between raw classification accuracy and computational feasibility for real-time embedded systems.

A. Performance Synthesis

As summarized in Table V, both supervised models significantly outperformed the unsupervised K-means baseline on

noisy data. While K-means reached an accuracy of 83% due to its inability to handle the complex overlap in a noisy feature space, supervised learning allowed the decision boundaries to adapt more precisely.

TABLE V
GLOBAL PERFORMANCE COMPARISON ON NOISY SIGNALS.

Model	Accuracy
K-means (Unsupervised)	83.00%
k-NN (Supervised)	90.75%
SVM (Supervised)	**90.50%**

The results demonstrate that the **SVM with an RBF kernel** is the most effective model for this task. Even if the SVM is a little bit under the k-NN in mean accuracy (only 0.25%), the SVM offers a significantly lower memory footprint during inference. Unlike k-NN, which requires storing the entire dataset, SVM only relies on a subset of critical points (support vectors). This efficiency, combined with its ability to define complex non-linear boundaries, proves its superior ability to generalize in non-stationary urban environments.

B. Architectural Suitability for ANC Systems

Beyond accuracy, the choice of a model for a Smart ANC trigger must consider hardware constraints:

1) **Memory Footprint:** k-NN require storing all the training feature vectors (MFCC and Energy Ratio) to perform classification. In contrast, the SVM only stores the support vectors, which represent a small fraction of the training set.
2) **Inference Latency:** For each new audio frame, k-NN must compute the Euclidean distance to every point in the database. The SVM inference simply requires evaluating the kernel function against the support vectors, offering a much lower and more predictable latency.
3) **Explainability:** Both models benefit from the physical relevance of our descriptors. The confusion matrices showed that errors are primarily linked to "speech-like" urban transients (horns/sirens) or segments with very low Signal-to-Noise Ratio (SNR).

C. Final Verdict

While k-NN provided a strong baseline, the **RBF-SVM** is the optimal choice for a transparency-mode trigger. Its lower memory requirements makes it the most viable candidate for integration into low-power wearable hardware.

V. COMPARISON WITH RELATED WORK

The challenge of Voice Activity Detection (VAD) and sound classification has been extensively studied, particularly within the DCASE community. It is essential to position our low-complexity approach against current state-of-the-art methodologies.

A. Deep Learning vs. Hand-crafted Features

Recent literature in audio pattern recognition is dominated by Deep Learning architectures. For instance, Convolutional Neural Networks (CNNs) operating on Log-Mel Spectrograms have achieved accuracies exceeding 95% on urban sound datasets [2]. However, these models often require millions of parameters and significant computational power. Our approach, utilizing only two discriminative descriptors (MFCC 2 variance and Energy Ratio), achieves a competitive 90.50% accuracy while being orders of magnitude lighter, making it suitable for the *always-on* constraints of ANC wearables.

VI. CONCLUSION

This study presented a robust framework for distinguishing human speech from urban noise, specifically designed for real-time integration into Smart ANC systems. By leveraging a dual-domain feature set consisting of MFCC standard deviation and Energy Ratio, we achieved a high degree of class separation with low computational complexity.

Our comparative analysis revealed that while unsupervised K-means provides a solid baseline for clean data, supervised models are essential to handle the spectral overlap of noisy environments. The SVM with an RBF kernel emerged as the optimal solution, achieving a mean accuracy of 90.50%. Beyond raw performance, the SVM's reliance on support vectors offers a significantly lower memory footprint and inference latency compared to k-NN, making it the most suitable candidate for battery-powered wearable hardware.

Future work could explore the integration of a temporal smoothing filter on the classifier's output to prevent rapid "mode switching" in ANC devices. Nevertheless, the current results validate that explainable, low-complexity descriptors combined with optimized SVMs provide a reliable trigger for situational awareness in urban acoustic monitoring.

REFERENCES

[1] A. Mesaros et al., "Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, 2018.

[2] J. P. Bello et al., "SONYC: A System for the Monitoring, Analysis and Mitigation of Urban Noise Pollution," *Comm. ACM*, vol. 62, no. 2, pp. 68–77, 2019.

[3] V. Panayotov et al., "Librispeech: an ASR corpus based on public domain audio books," in *Proc. IEEE ICASSP*, 2015.

[4] B. McFee et al., "librosa: Audio and music signal analysis in python," in *Proc. 14th Python in Science Conf.*, 2015.