

Voicemail Detection in Telephone Recordings for Call/Contact Center Systems Using Hierarchical Spectral Windowing

Michał Zawadzki
0009-0003-7438-6022
Kielce University of Technology
Faculty of Electrical Engineering,
Automatic Control and Computer
Science,
Kielce, Poland
m.zawadzki@tu.kielce.pl

Mirosław Płaza
0000-0001-9728-3630
Kielce University of Technology
Faculty of Electrical Engineering,
Automatic Control and Computer
Science,
Kielce, Poland
m.plaza@tu.kielce.pl

Marcin Kowalczyk
0000-0001-6137-0580
Warsaw University of Technology
Faculty of Electronics
and Information
Technology,
Warsaw, Poland
marcin.kowalczyk@pw.edu.pl

Abstract—The paper proposes a custom system for automatic voicemail detection designed for production-grade call/contact center environments. The detection problem is formulated as a binary classification task, where the objective is to distinguish live speech from a recorded voicemail message. The proposed system is based on a two-level hierarchical audio signal windowing structure combined with dominant spectral component extraction using the Discrete Fourier Transform. The extracted feature representation serves as input to a multi-layer perceptron performing binary classification. The system was designed for telephone recordings sampled at eight kHz. Experiments were conducted on a dataset of 7.321 authentic telephone recordings from a production contact center. A systematic search over filtering threshold parameters yielded 56 experiments. The best accuracy of 93.35% was achieved using a small window threshold of 1.5 and a large window threshold of 0.1. The results confirmed the effectiveness of the proposed approach and highlighted the role of filtering parameters in classification accuracy.

Index Terms—voicemail detection, call/contact center, Discrete Fourier Transform, spectral feature extraction, hierarchical windowing, multilayer perceptron.

I. INTRODUCTION

IN RECENT years, interpersonal communication and customer service have increasingly relied on various computer-based systems. A particularly prominent role in this context is played by call/contact center (CC) systems, whose presence is strongly visible across numerous business domains, including marketing, sales, and debt collection [1]. These systems are also gaining increasing significance in the medical domain [2].

The highly competitive market in which these systems operate demands their continuous development. Advances are evident across many areas, including automatic speech transcription methods [3], voicebots and chatbots [4], emotion recognition methods dedicated to CC systems [5], [6], and multi-label classification of text messages [7], [8].

In CC systems, outbound traffic plays an important role [9], as its efficiency directly translates into key performance indicators (KPIs) such as the number of successfully completed calls, customer satisfaction levels, and the cost per contact [10], [11]. In many outbound campaigns, particularly those related to debt collection activities, a significant operational challenge arises from calls going unanswered by the called party. In such cases, the call is forwarded to a voicemail system, effectively making the contact attempt resource-intensive but inefficient [12]. Therefore, its essential to detect as early as possible that the call has been answered not by a live person, but by a voicemail system. The ability to distinguish between a live caller and a recorded voicemail greeting determines how the system handles the ongoing call. The call may either be transferred to a consultant for a live conversation or discarded as voicemail contact. This has a direct impact on consultant efficiency and service quality. The detection of Interactive Voice Response (IVR) messages and voicemail greetings goes beyond classical Voice Activity Detection (VAD), requiring dedicated methods capable of distinguishing voicemail messages from genuine customer contact [13]. Such solutions are referred to as Answering Machine Detection (AMD), and their presence becomes critical in these scenarios.

Despite the practical significance of this issue, voicemail detection has not yet received sufficient attention in scientific literature. The number of publicly documented algorithms is limited, and the solutions deployed by CC system vendors in production environments are generally not disclosed [14]. An additional constraint is the nature of quality audio using in call centers which are standardly sampled at eight kHz. This significantly reduces the available spectral resolution and renders speech processing techniques developed for high-quality audio suboptimal in this environment.

The primary contribution of this work is the proposal of a custom two-level hierarchical audio signal windowing structure as the basis for feature representation in a binary classification task. While signal analysis is a well-studied field, the application of two-level temporal windowing as a direct input representation for a Multi-Layer Perceptron (MLP) classifier in the context of voicemail detection in telephone recordings has not been previously investigated. An additional contribution is the empirical validation of the proposed approach on a real production dataset and the analysis of the influence of filtering parameters on classification quality.

The rest of this paper is organized as follows. Section II reviews existing approaches to voicemail detection. Section III describes the proposed research methodology. Section IV presents the obtained results. Section V contains a discussion of the results and a description of the method's limitations. Section VI provides conclusions and directions for future work.

II. RELATED WORKS

Existing approaches to voicemail detection can be divided into three main categories: acoustic methods based on the detection of a characteristic audio tone, methods exploiting metadata and signaling information available in telecommunication protocols, and machine learning-based methods that analyze properties of the audio stream. Each of these approaches operates at a different processing level and involves a distinct trade-off between decision latency, implementation complexity, and robustness to variability in the production environment.

The traditional approach is the detection of the characteristic beep tone generated by voicemail systems at the end of the greeting message, signaling the transition to recording mode. From an acoustic standpoint, this signal is most commonly a sinusoidal tone of constant frequency, whose spectral density is characterized by a distinct energy peak, clearly distinguishing it from speech, which is time-varying in nature. Frequency-domain analysis methods such as the Discrete Fourier Transform or the Goertzel algorithm are used for its identification [15]–[17]. This method is characterized by low computational complexity and high precision; however, its fundamental limitation is decision latency. Detection requires waiting for the entire greeting message to complete, which may take from several to over ten seconds, significantly reducing the operational efficiency of outbound campaigns. Furthermore, not all voicemail systems generate a beep tone, and its parameters may vary depending on the operator or PBX configuration.

An alternative is provided by methods based on Session Initiation Protocol (SIP) signaling metadata, which under favorable conditions allows a classification decision to be made almost instantaneously, without requiring any audio stream analysis [18]. The simplest indicator is the call answer time: voicemail systems answer calls programmatically, resulting in response times close to zero, whereas a live caller requires

time for a physical reaction. Additional information is provided by SIP headers such as Diversion (RFC 5806) and History-Info (RFC 7044), which may directly indicate that the call has been forwarded to a voicemail system [19]–[21]. However, this method is strongly dependent on the telecommunication operator's policy. Not all operators transmit complete signaling information, and in environments with incomplete SIP chains, such data may be partially or entirely unavailable. Furthermore, a fast answer time may result from a user's auto-answer feature, potentially leading to false classifications. This method is therefore effective only as a complement to other approaches.

The broadest classification potential is offered by machine learning-based methods. Unlike beep tone detection, which requires waiting for the entire greeting to end, these methods are capable of making a decision based on analysis of only the first few seconds of the audio stream, thereby eliminating the need to wait for a characteristic tone. This constitutes a significant operational advantage in CC environments. The literature documents the use of Mel-Frequency Cepstral Coefficients (MFCC) [14], [22], the GeMAPS and eGeMAPS parameter sets [13], [23], and deep feature vectors generated by the YAMNet model [24] as input representations for binary classifiers. On the classifier architecture side, both Convolutional Neural Networks (CNN) operating on spectrograms [25] and recurrent neural networks with GRU layers processing sequences of feature vectors [14] have been adopted. These methods demonstrate high effectiveness. However, their primary drawback is the significant cost of development. They require large, manually labeled training datasets that are difficult to obtain due to the private nature of telephone recordings and legal constraints related to personal data protection [13]. A further limitation is the generalization capability of trained models. Classifiers trained on data from one operator may exhibit a significant drop in performance when deployed in a different environment.

The two-level windowing structure proposed in this work fundamentally alters the processing logic. It enables extraction of both local and global spectro-temporal signal properties without relying on external feature extraction libraries. Its key advantage over other approaches lies in the way the incoming audio signal is processed. Classical voicemail detection methods treat a recording as a static, finite data object to be processed. This applies to both acoustic methods and machine learning-based approaches. Beep detection requires waiting for the entire greeting to end, which may take several seconds. Methods based on features such as MFCC or GeMAPS also typically operate on the full recording or a predetermined fixed-length segment, meaning the classification decision can only be made after a required number of samples has been collected. In both cases, the system remains passive throughout the data acquisition phase, and classification occurs as a one-time event upon completion of this process.

The proposed two-level windowing structure fundamentally changes this processing logic. Since each large window is constructed, processed and classified independently of the

others, the system is capable of making classification decisions continuously, as successive signal samples arrive, even while the telephone call is still in progress. There is no single global time threshold after which the system makes its decision. Instead, each large successive window provides an independent assessment that can be aggregated or updated in real time. The decision threshold is therefore dynamic and adaptive in nature: the more windows that have been classified, the greater the confidence in the final decision. The practical implications of this approach are significant from the perspective of a CC production environment. For short voicemail messages, the system may reach the correct decision after processing only the first few large windows, without needing to process the entire recording. For longer calls, each successive window provides additional information, gradually improving classification confidence. The system therefore does not need to assume a maximum recording length as a prerequisite for decision-making - it is sufficient that enough samples have arrived to fill a single large window. This makes the proposed approach inherently better suited to streaming environments than methods requiring analysis of the entire recording as a single unit.

III. RESEARCH METHODOLOGY

This section describes the proposed research methodology. Section A characterizes the dataset used in the experiments. Sections B through G describe the successive stages of the proposed processing pipeline, from initial audio signal pre-processing, through hierarchical windowing and spectral feature extraction, to the classifier architecture.

A. Dataset

The neural network model was trained and validated on an audio dataset consisting of 7.321 recordings in WAV format. The dataset originates from a real telecommunication environment and represents authentic telephone calls recorded during outbound call center campaigns.

TABLE 1. DATASET STATISTICS

Feature	Value
Total number of recordings	7.321
Class 0 - live speech	3.404 (46,5%)
Class 1 - voicemail	3.917 (53,5%)
File format	WAV
Sampling frequency	8 kHz
Number of channels	2 (stereo)
Mean recording duration	~8 s
Mean file size	~200 KB
Training set (80%)	5.856 recordings
Test set (20%)	1.465 recordings
Collection periods	IX 2022, X 2022, I 2024

Data was collected across three time periods: September 2022, October 2022, and January 2024, ensuring the representativeness of the dataset with respect to seasonal variability and the evolution of voicemail systems over time. The dataset was divided into two classes: Class 0, corresponding to recordings containing live caller speech (3.404 recordings, 46.5%), and Class 1, corresponding to recordings containing a voicemail message (3.917 recordings, 53.5%). The slight class imbalance present in the raw recordings was addressed during training by applying random undersampling of the majority class. Each recording was manually labeled by telecommunications operators, ensuring high label quality. Detailed dataset statistics are presented in Table 1.

All files were recorded in stereo, with one channel capturing the agent's voice and the other capturing the caller's voice or the voicemail message. Due to the short duration of the recordings (approximately eight seconds), the material does not contain sensitive personal data. It should be noted that the unit of classification in the proposed system is not the entire recording, but a single large time window generated during the hierarchical windowing process. Each recording yields multiple such windows, resulting in a total number of feature vectors in the training set on the order of 100.000, despite the base dataset comprising just over 7.000 files.

B. System Architecture

The proposed system is designed as a processing pipeline consisting of five cascaded stages:

- Audio signal preprocessing - extraction of the relevant channel from the stereo recording and signal length normalization.
- Hierarchical temporal windowing - segmentation of the signal into a two-level window structure comprising small local windows and large contextual windows, each grouping a fixed number of small windows.
- Spectral feature extraction via DFT - computation of dominant spectral components for each small window and their concatenation into a feature vector representing the entire large window.
- Threshold-based filtering with normalization - elimination of silence and noise segments based on a two-stage energy criterion, where at the small window level the total energy (amplitude) of dominant spectral components is evaluated, and at the large window level the fraction of windows classified as active is verified. Feature values are subsequently scaled to the range [0, 1].
- Binary classification using a multilayer neural network - assignment of the feature vector to one of two classes: live caller speech or voicemail message.

Each stage is tightly coupled to the next, the output of one stage constitutes the input to the following stage. The system

features a modular architecture, enabling independent configuration of individual components without requiring reconstruction of the entire pipeline.

C. Preprocessing

The system input consists of WAV recordings sampled at eight kHz. For two-channel recordings, the system extracts a single channel, and all subsequent processing is performed on the resulting mono signal. This operation is motivated by the specific structure of the available telephone recordings, in which one channel carries the caller's voice or the voicemail message. After channel extraction, the signal is either trimmed or zero-padded to a fixed number of samples. Zero-padding ensures a uniform input length across all recordings, regardless of their actual duration. This is a necessary condition for data vectorization, as multilayer neural networks require inputs of fixed dimensionality.

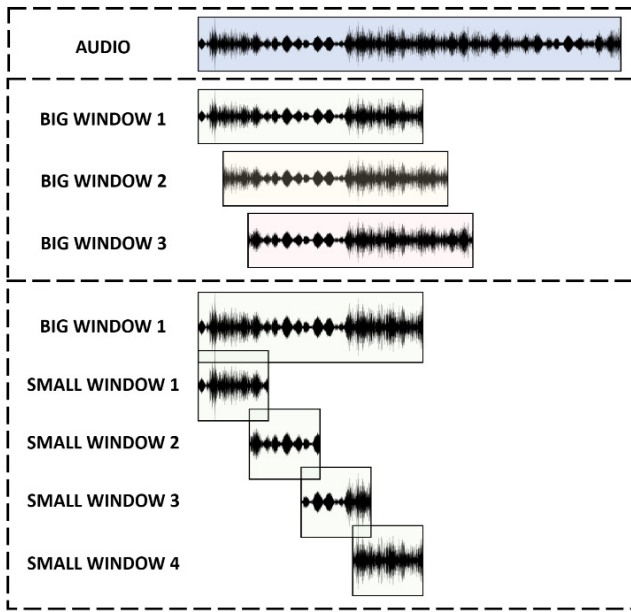


Fig 1. Schematic of the two-level hierarchical windowing structure.

D. Hierarchical Temporal Windowing

Direct analysis of an entire recording as a single unit is insufficient for capturing the temporal variability of the speech signal. Short segments of a recording may differ substantially in their spectral characteristics. For this reason, the system employs a two-level windowing structure, which constitutes the primary contribution of this work and enables signal analysis at both local and global scales.

At the lowest level of the hierarchy, small time windows of fixed length are defined. The default value is 256 samples, which at a sampling frequency of eight kHz corresponds to a time interval of 32 ms - a range widely used in speech analysis that approximates the duration of a phoneme [26], [27]. Successive small windows are shifted by a configurable number of samples. The system supports two basic configurations: a shift equal to the window length (non-overlapping) and a shift

equal to a fraction of the window length (overlapping), which allows features at the boundary between adjacent segments to be captured.

At the higher level of the hierarchy, the signal is segmented into large windows, each encompassing a fixed number of small windows. This structure allows the classifier to consider both the local spectral properties of individual segments and the broader temporal context spanning several hundred milliseconds of the recording. This is particularly relevant for distinguishing spontaneous speech from recorded voicemail messages, as these two signal types differ not only at the phoneme level but also in the overall structure of the utterance. The two-level windowing scheme is illustrated in Fig. 1.

The baseline windowing parameter configuration used in the experiments is presented in Table 2.

TABLE 2. BASELINE WINDOWING PARAMETER CONFIGURATION

Parameter	Value	Description
N_LEN_SMALL_WINDOW	256	Samples small window length
N_SHIFT_SMALL_WINDOW	256	Samples small window shift
N_SMALL_WINDOW	50	Number of small windows per large window
N_BIG_WINDOW	50	Number of large windows
N_SHIFT_BIG_WINDOW	128	Samples large window shift

E. Spectral Feature Extraction

Spectral feature extraction is the central stage of the system. Its purpose is to transform the raw time-domain signal into a compact numerical representation describing the frequency content of each segment. The adopted approach is based on the DFT. Rather than storing the full amplitude spectrum, only the dominant components with the highest amplitudes are selected from each segment, significantly reducing the dimensionality of the representation while retaining the most energetically significant components of the signal [28].

Processing is performed independently for each small window as follows. Prior to computing the transform, each signal segment is multiplied by a window function to eliminate spectral leakage arising from discontinuities at the segment boundaries. The Hamming window is applied, defined as:

$$w[n] = 0,54 - 0,46 \cdot \cos\left(\frac{2\pi n}{N-1}\right), n = 0, 1, \dots, N-1$$

where N denotes the number of samples in the small window. The choice of the Hamming window is motivated by its widespread use in speech signal processing and its favorable spectral leakage suppression characteristics [29], [27]. Based on the element-wise product of the signal and the window function, the amplitude spectrum is computed, and the dominant frequency components are selected, as described in detail below.

An N -point DFT is computed on the windowed signal. Since the input signal is real-valued, the DFT spectrum is symmetric about the Nyquist frequency [30]. Therefore, only the first half of the spectrum is retained, corresponding to frequencies from 0 Hz to 4 kHz. The real part of the transform output is extracted, and its magnitude is computed to yield a simplified amplitude spectrum. Prior to component selection, the spectrum is trimmed to a parameterized frequency band, eliminating components unrelated to the acoustic content of the call. From the band-limited spectrum, a fixed number of indices with the highest amplitudes are selected and converted to frequency values. Each frequency and amplitude pair constitutes one element of the small window feature representation. The feature extraction parameters are summarized in Table 3.

TABLE 3. SPECTRAL FEATURE EXTRACTION PARAMETERS

Parameter	Value	Description
N_MAX	5	Number of dominant components
MIN_HZ	50 Hz	Lower bound of analysis band
MAX_HZ	1.800 Hz	Upper bound of analysis band
Small window vector dimension	10	$N_MAX * 2$
Large window vector dimension	500	$N_SMALL_WINDOW * NMAX * 2$

F. Threshold-Based Filtering and Normalization

Raw feature vectors contain both active speech segments and fragments of silence or background noise. To eliminate these components, a two-stage threshold-based filtering procedure is applied. At the small window level, each segment is evaluated based on a logarithmic energy measure of its dominant spectral components. If this measure falls below a pre-defined threshold, the segment is zeroed and treated as silence. At the large window level, the fraction of non-zero small windows -are computed. If it falls below a second threshold, the entire large window is discarded. Following filtering, normalization is applied to both frequency and amplitude values, mapping all values to the range [0, 1]. The filtered data are subsequently flattened into a one-dimensional feature vector that serves as the classifier input. The system supports three feature vector configurations: frequencies only, amplitudes only, or both types interleaved. This last configuration was adapted as the default.

G. Classifier Architecture

The feature vector is fed into an MLP constructed as a sequential model. All hidden layers employ the Rectified Linear Unit (ReLU) activation function. The output layer consists of a single neuron with a sigmoid activation function, producing a probability score for class membership in the voicemail class. Binary cross-entropy is used as the loss function, and the Adam optimizer is employed for training.

IV. RESULTS

The experiments were designed to investigate the effect of filtering threshold parameters on binary classification quality in the proposed system. Two filtering thresholds were selected as variables: the small window energy threshold and the large window percentage threshold. All remaining system parameters, including the MLP classifier architecture, windowing parameters, and feature extraction parameters, were held constant throughout the experiments. Their values were determined experimentally through preliminary investigations and guided by the authors' experience with MLP-based classifiers to provide a stable operating point for the proposed method. The selected values are listed in Table 4.

TABLE 4. PARAMETERS HELD CONSTANT ACROSS ALL EXPERIMENTS

Parameter	Value
N_LEN_SMALL_WINDOW	256 (32 ms)
N_SHIFT_SMALL_WINDOW	256
N_SHIFT_BIG_WINDOW	128
N_MAX	5
N_SMALL_WINDOW	50
N_BIG_WINDOW	50
MIN_HZ	50 Hz
MAX_HZ	1.800 Hz
Feature vector type	Frequencies and amplitudes
Classifier architecture	MLP
Number of epochs	50
Batch size	32

For the small window energy threshold, eight values were evaluated: 0, 0.1, 0.3, 0.7, 1.0, 1.5, 2.0, and 2.5. For the large window percentage threshold, seven values were evaluated: 0, 0.1, 0.2, 0.3, 0.4, 0.6, and 0.8. The full combinatorial grid yields 56 experiments in total. Each experiment encompassed the complete processing pipeline: feature extraction, threshold-based filtering, normalization, dataset balancing, MLP classifier training, and evaluation on the test set. Accuracy on the test set was adopted as the primary performance metric. This choice was made to enable direct comparison with previously published studies, in which classification performance is commonly reported in terms of accuracy.

Of the 56 experiments carried out, a subset resulted in model degeneration, manifested as an accuracy value oscillating around 0.5 - a characteristic symptom of a classifier adopting a constant-class prediction strategy. This phenomenon occurred most frequently at low values of both thresholds, when the feature vectors contained a substantial proportion of silence and noise segments. Experiments in which the model learned to classify effectively achieved test set accuracy values ranging from approximately 0.87 to 0.93. Given the simplicity of the adopted feature representation and the narrow-band nature of the input data, these results can be considered

promising. A comprehensive overview of the results for all threshold combinations is presented as a heatmap in Fig. 2. The values are reported as percentages with one decimal place. Gray cells correspond to configurations in which model degeneration occurred.

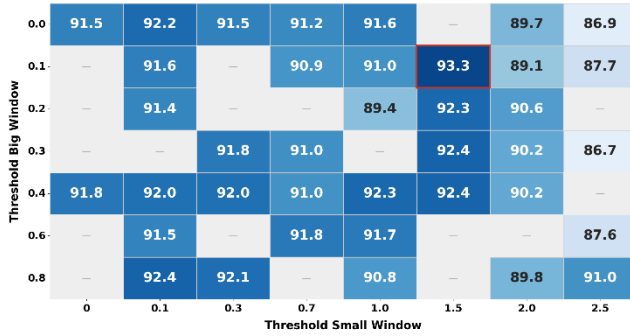


Fig. 2. Accuracy (%) as a function of the small window threshold and the large window threshold. Gray cells indicate model degeneration. The best result is highlighted with a red border.

Analysis of the heatmap reveals a clear pattern associated with a small window threshold of 1.5. In all configurations in which the model achieved effective classification, accuracy exceeded 92%. This result is consistently higher than those obtained for all other values of this parameter, where accuracy varied over a wider range and model degeneration occurred more frequently. This indicates that a threshold of 1.5 constitutes a particularly favorable operating point for the dataset under study - it effectively eliminates low-energy segments while retaining a sufficient number of active speech segments for effective classifier training.

To further investigate the effect of the large window threshold, a detailed analysis was conducted for a fixed small window threshold of 1.5. Table 5 presents the relationship between the large window threshold value, the size of the training set, the training time, and the resulting accuracy. The value of Nan in the last column indicates model degeneration.

The data in Table 5 reveal two notable trends. Firstly, as the large window threshold increases, the number of training samples decreases monotonically, from 98.150 at a threshold of 0.0 to 65.904 at a threshold of 0.8. The accompanying reduction in training time is nearly proportional, indicating a linear relationship between dataset size and the computational cost of the training process. This relationship is illustrated in Fig. 3. Secondly, the best accuracy (93.35%) was achieved at a threshold of 0.1 rather than 0.0, despite the second configuration having a larger training set. This suggests that even minimal filtering at the large window level improves the quality of the feature representation to a degree that outweighs the benefit of having more training examples. To provide a more comprehensive evaluation of the best performing configuration, additional metrics were also calculated. The model achieved a precision of 97.28%, a recall of 90.87%, and an F1-score of 93.97%, confirming that the high accuracy is accompanied by a well-balanced classification performance.

TABLE 5. EFFECT OF THE LARGE WINDOW THRESHOLD ON TRAINING SET SIZE, TRAINING TIME, AND ACCURACY FOR SMALL WINDOW THRESHOLD = 1.5

Large window threshold	Training samples	Training time (s)	Accuracy
0	98.150	1.668	Nan
0,1	95.508	1.643	93,35%
0,2	93.383	1.606	92,28%
0,3	90.812	1.590	92,43%
0,4	87.796	1.552	92,39%
0,6	79.675	1.464	Nan
0,8	65.904	1.306	Nan

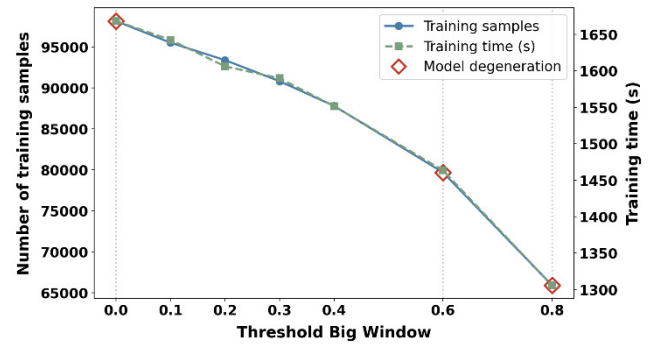


Fig. 3. Number of training samples (left axis) and training time in seconds (right axis) as a function of the large window threshold for small window threshold = 1.5.

The analysis is complemented by Fig. 4, which presents the distribution of accuracy values for each evaluated small window threshold value. Each bar corresponds to a single non-degenerate experiment. The visualization confirms the observation regarding threshold 1.5. All experiments carried out at this parameter value cluster above the 92% line, whereas for all other threshold values the spread of results is considerably wider and a portion of experiments falls below 90%.

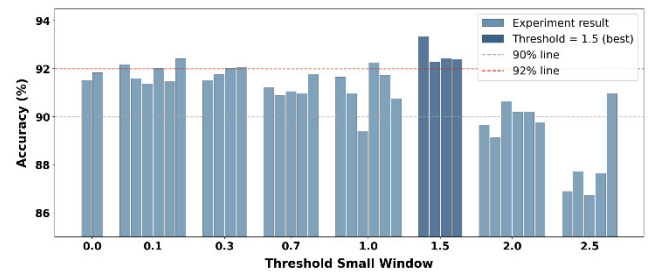


Fig. 4. Distribution of accuracy values for each evaluated small window threshold value.

V. DISCUSSION AND LIMITATIONS

The obtained results confirm that the proposed feature representation, based on dominant spectral components within a

two-level windowing structure, is capable of effectively distinguishing live speech from recorded voicemail messages. The best accuracy of 93.35% was achieved with a relatively simple classifier architecture, indicating that the selected spectral components carry high-quality discriminative information. The obtained accuracy is comparable to the results reported in the literature. Previous studies achieved accuracies of approximately 94%, 93%, 96%, and 94% [12]–[14], [25]. However, these methods were evaluated using considerably longer speech segments, whereas the proposed approach operates on much shorter audio fragments. The comparable performance obtained on substantially shorter recordings suggests that the proposed method is a promising direction for practical call center applications. A particularly noteworthy observation is the stability of results for a small window threshold of 1.5. While other values of this parameter led to a wide spread of results, a threshold of 1.5 consistently yielded accuracy above 92% across all non-degenerate configurations. Such stability is a highly desirable property in a production environment, where the predictability of system behavior is as important as peak performance.

The experiments covered only a search over the filtering threshold parameter space, with all other system parameters kept constant. The impact of window size, window overlapping, and the number of dominant spectral components remains uninvestigated. The irregular pattern of model degeneration across the parameter space further suggests that some results may be sensitive to random weight initialization, which requires verification through repeated experiments with averaged results. Finally, the dataset was collected from a single call center operator, limiting the ability to evaluate the model's portability to other telecommunication environments. Although the recordings were collected over different time periods, which introduced a degree of natural variability, all data originated from the same operational environment. Acquiring similar datasets from multiple operators is challenging because call recordings are highly sensitive and subject to strict privacy and regulatory constraints. Therefore, an important direction for future work is the evaluation of the proposed method on independent external datasets collected from different call center operators to assess its generalization capability. All of the aforementioned limitations will be addressed in future work aimed at further developing the proposed methodology.

VI. CONCLUSION

This paper proposes a voicemail detection system based on a two-level hierarchical audio signal windowing structure and dominant spectral component extraction via the DFT. Binary classification is performed by a multilayer perceptron. The system was designed for production-grade call center environments in which recordings are sampled at eight kHz. A total of 56 experiments were conducted through a systematic search over the filtering threshold parameter space. The best accuracy of 93.35% was achieved at a small window threshold of 1.5 and a large window threshold of 0.1.

The results confirm that a simple representation based on dominant spectral components, appropriately filtered and normalized, provides the classifier with sufficient information for effective classification. Among all evaluated threshold values, a small window threshold of 1.5 distinguished itself by exceptional stability. Accuracy exceeded 92% in all non-degenerate configurations. Such consistent system behavior is a particularly desirable property in a production environment, where operational reliability is as critical as peak classification performance.

A key architectural property of the proposed system is its natural suitability for streaming processing. Each large window is constructed, processed, and classified independently, enabling classification decisions to be made continuously as signal samples arrive. The system does not require waiting for the recording to end. A decision can be reached as soon as the first large windows have been collected. This constitutes an operational advantage over methods that treat the recording as a closed unit of analysis.

Future work should include an expanded set of evaluation metrics and confusion matrix analysis. A full parameter space search covering window size, window overlap, and the number of dominant spectral components is warranted. A promising direction is the comparison of the proposed MLP classifier with sequential architectures such as recurrent neural networks with GRU or LSTM layers, which may more effectively exploit the temporal structure of hierarchical feature vectors. Finally, the system requires validation on recordings from multiple telecommunication operators to assess its ability to generalize beyond the environment in which the training data was collected.

ACKNOWLEDGMENT

During the preparation of this work, the authors used Claude (Anthropic) for language editing, phrasing suggestions, and limited drafting assistance. The authors take full responsibility for the content of this publication.

REFERENCES

- [1] C. Sánchez, S. Maldonado and C. Vairetti, Improving debt collection via contact center information: A predictive analytics framework. *Decision Support Systems*, 2022. doi.org/10.1016/j.dss.2022.113812.
- [2] M.H. Rim, K.C. Thomas, J. Chandramouli, S.A. Barrus and N.A. Nickman, Implementation and quality assessment of a pharmacy services call center for outpatient pharmacies and specialty pharmacy services in an academic health system. *Am J Health Syst Pharm*. 2018. doi: 10.2146/ajhp170319.
- [3] A. Graves, A. Mohamed and G. Hinton, Speech recognition with deep recurrent neural networks, 2013 IEEE International Conference on Acoustics. Speech and Signal Processing, Vancouver, BC, Canada, 2013, pp. 6645-6649. doi: 10.1109/ICASSP.2013.6638947.
- [4] T. Noga, The use of chatbots and voicebots by public institutions in the communication process with clients. *Zeszyty Naukowe. Organizacja i Zarządzanie/Politechnika Śląska*, 2023. doi:10.29119/1641-3466.2023.174.6.
- [5] M. Płaza, R. Kazała, Z. Koruba, M. Kozłowski, M. Lucińska, K. Sitek, and J. Spyłka, Emotion Recognition Method for Call/Contact Centre Systems. *Applied Sciences*, vol. 12, no. 21, p. 10951, 2022. doi.org/10.3390/app122110951.

- [6] M. Bojanić, V. Delić and A. Karpov, Call Redistribution for a Call Center Based on Speech Emotion Recognition. *Applied Sciences*. 2020; 10(13):4653. doi.org/10.3390/app10134653.
- [7] K. Poczęta, M. Plaza, T. Michno, M. Krechowicz, and M. Zawadzki, A multi-label text message classification method designed for applications in call/contact centre systems, *Applied Soft Computing*, vol. 145, pp. 1–20, 2023. doi.org/10.1016/j.asoc.2023.110562.
- [8] Z. Ahanin, M.A. Ismail, N.S.S. Singh and A. AL-Ashmorim, Hybrid Feature Extraction for Multi-Label Emotion Classification in English Text Messages. *Sustainability*. 2023; 15(16):12539. doi.org/10.3390/su151612539.
- [9] M. Kozłowski, M. Plaza, M. Lucińska and M. Kowalczyk, Metoda predykcji ruchu wychodzącego w systemach klasy call/contact center wykorzystująca algorytm uczenia ze wzmocnieniem, *Przegląd Telekomunikacyjny - Wiadomości Telekomunikacyjne*, vol. 4, pp. 348–351, 2025. doi:10.15199/59.2025.4.78.
- [10] M. Plaza and Ł. Pawlik, Influence of the Contact Center Systems Development on Key Performance Indicators, *IEEE Access*, vol. 9, pp. 44580–44591, 2021. doi: 10.1109/ACCESS.2021.3066801.
- [11] C.S.C Correia, Optimizing Call-Center Operations with Reinforcement Learning. MS thesis. Universidade do Porto (Portugal), 2020.
- [12] E. Minaj, Optimization of answering machine detection application in asterisk. 2020. lutpub.lut.fi/handle/10024/161698.
- [13] A. Kopylov, O. Seredin, A. Filin and B. Tyshkevich, Detection of interactive voice response (IVR) in phone call records. *International Journal of Speech Technology* 23.4 2020: 907-915. doi.org/10.1007/s10772-020-09754-3.
- [14] K. Altwlkany, S. Delalić, E. Selmanović, A. Alihodžić and I. Lovrić, A Recurrent Neural Network Approach to the Answering Machine Detection Problem. 2024 47th MIPRO ICT and Electronics Convention (MIPRO). IEEE, 2024. doi: 10.1109/MIPRO60963.2024.10569812.
- [15] T. Joseph, K. Tyagi and R. Kumbhare, Quantitative analysis of DTMF tone detection using DFT, FFT and Goertzel algorithm. 2019 Global Conference for Advancement in Technology (GCAT). IEEE, 2019. doi: 10.1109/GCAT47503.2019.8978284.
- [16] V. Balambica, N. Hamadamen, A. Karthikayen and M. Praveen, Digital signal processing dual tone multifrequency detector. 2023. doi:10.37896/YMER22.02/98.
- [17] P. Sysel and P. Rajmic, Goertzel algorithm generalized to non-integer multiples of fundamental frequency. *EURASIP J. Adv. Signal Process.* 2012, 56 (2012). https://doi.org/10.1186/1687-6180-2012-56.
- [18] D. Pereira and R. Oliveira, Detection of Abnormal SIP Signaling Patterns: A Deep Learning Comparison. *Computers* 2022. doi.org/10.3390/computers11020027.
- [19] H. Luo and Y. Zhang, Beep tone detection within RTP streams based on TK energy operator and DESA2 algorithm, 2011 International Conference on Electronics, Communications and Control (ICECC), Ningbo, China, 2011, pp. 766-769, doi: 10.1109/ICECC.2011.6066454.
- [20] https://www.rfc-editor.org/rfc/rfc7044.txt.
- [21] https://datatracker.ietf.org/doc/html/rfc4458.
- [22] J.K. Deters, S. Janus, J.A.L. Silva, H.J. Wörtche and S.U. Zuidema, Sensor-based agitation prediction in institutionalized people with dementia A systematic review, *Pervasive and Mobile Computing*, 2024. doi.org/10.1016/j.pmcj.2024.101876.
- [23] B.T. Atmaja and M. Akagi, On The Differences Between Song and Speech Emotion Recognition: Effect of Feature Sets, Feature Types, and Classifiers, 2020 IEEE REGION 10 CONFERENCE (TENCON), Osaka, Japan, 2020, pp. 968-972, doi: 10.1109/TENCON50793.2020.9293852.
- [24] S. Lachenani, H. Kheddar and M. Ouldzmilri, Improving Pretrained YAMNet for Enhanced Speech Command Detection via Transfer Learning, 2024 International Conference on Telecommunications and Intelligent Systems (ICTIS), Djelfa, Algeria, 2024, pp. 1-6, doi: 10.1109/ICTIS62692.2024.10894266.
- [25] N. Anisimov, M. Olkhovskiy and K. Kishinsky, Answering machines detection in contact centers using convolutional neural networks. url: www.fiztech-usa.net/anisimov/papers/answering-machines-detection_6.pdf.
- [26] B. Jiang and J. Yang, Preferred frame length for the short-time magnitude spectrum on speech intelligibility and speech quality, 2011 8th International Conference on Information, Communications & Signal Processing, Singapore, 2011, pp. 1-3, doi: 10.1109/ICICS.2011.6174266.
- [27] G. Ma, P. Hu, J. Kang, S. Huang and H. Huang, Leveraging Phone Mask Training for Phonetic-Reduction-Robust E2E Uyghur Speech Recognition, 2022. doi.org/10.21437/Interspeech.2021-964.
- [28] A. B. Abdusalomov, F. Safarov, M. Rakhimov, B. Turaev and T.K. Whangbo, Improved Feature Parameter Extraction from Speech Signals Using Machine Learning Algorithm. *Sensors* 2022, 22, 8122. doi.org/10.3390/s22218122.
- [29] M. Pereira, S. Chapaneri and D. Jayaswal, Analysis of windowing techniques for speech emotion recognition, 2016 International Conference on Information Communication and Embedded Systems (ICES), Chennai, India, 2016, pp. 1-6, doi: 10.1109/ICES.2016.7518859.
- [30] J. G. Proakis and D.G. Manolakis, Digital Signal Processing: Principles, Algorithms, and Applications, Prentice-hall international INC. 1996. ISBN-10: 0133942899.