BULLETIN of Cherkasy State Technological University

Journal homepage: https://bulletin-chstu.com.ua/en

Vol. 30 No. 4. 2025

UDC 0004.02
DOI: 10.62660/bcstu/4.2025.119

A method for keyword recognition in voice signals in resource-constrained computer systems

Andrii Didus*
Postgraduate Student
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
03056, 37 Beresteyskyi Ave., Kyiv, Ukraine
https://orcid.org/0009-0004-2235-6742

Ihor Tereikovskyi
Doctor of Technical Sciences, Professor
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
03056, 37 Beresteyskyi Ave., Kyiv, Ukraine
https://orcid.org/0000-0003-4621-9668

Abstract. Keyword spotting on embedded platforms must balance accuracy and strict resource limits while remaining independent of network connectivity. The aim of the study was to develop and experimentally validate a classical, frugal recognition method that increases feature informativeness without increasing model complexity and is suitable for autonomous use on edge devices that rely only on a central processing unit. A weighted acoustic fingerprinting mechanism was proposed. Mel-frequency cepstral coefficients, together with their derivatives, were reweighted, aggregated, and serialised into compact discrete "fingerprints", which were then classified using the Levenshtein edit distance. Experiments were carried out on a Ukrainian-language command corpus from six native speakers (three male, three female), recorded with both headsets and far-field microphones; lexicons of 10, 100, and 200 words were evaluated under speaker-independent splits of 70%/15%/15%. The methodology comprised fixed parametrisation of mel-frequency cepstral coefficients, construction of a static weighting vector, voice-activity detection with spectral subtraction, uniform quantisation and serialisation, and deterministic edit-distance classification; for comparison, equal-weight baselines, hidden Markov models with Gaussian mixture emissions, Dynamic Time Warping, a lightweight convolutional neural network, and a reference depthwise-separable convolutional neural network were considered. The proposed method achieved macro-averaged harmonic means of precision and recall of 0.96/0.92/0.89 for 10/100/200-word lexicons in clean audio, and 0.78 at a signal-to-noise ratio of 5 decibels (100-word lexicon). The implementation required approximately 250 kilobytes of memory and operated with a real-time factor of 0.005 on Raspberry Pi 4 with 4 gigabytes, i.e., faster than real time. Superiority over equal-weight baselines, hidden Markov models with Gaussian mixture emissions, and Dynamic Time Warping was demonstrated, with performance approaching that of a compact convolutional neural network. It is concluded that weighted acoustic fingerprinting provides a robust, efficient, and autonomous keyword-spotting solution for deployments that use only a central processing unit

Keywords: embedded edge computing; acoustic fingerprinting; feature reweighting; edit-distance-based classification; robust speech commands; resource-constrained devices

INTRODUCTION

Effective keyword spotting (KWS) was a cornerstone technology for modern human-machine interfaces, particularly within the then-growing domain of autonomous and embedded systems, such as unmanned ground vehicles and smart home devices. The primary relevance of this task stemmed from the critical need for

Article's History: Received: 17.06.2025; Revised: 20.10.2025; Accepted: 15.12.2025.

Suggested Citation:
Didus, O., & Tereikovskyi, I. (2025). A method for keyword recognition in voice signals in resource-constrained computer systems. *Bulletin of Cherkasy State Technological University*, 30(4), 119-127. doi: 10.62660/bcstu/4.2025.119.

*Corresponding author

reliable and low-latency voice-command activation in environments where computational power, energy consumption, and network connectivity were severely restricted. This technological challenge established a significant scientific problem: the development of recognition methods that achieved an optimal equilibrium between classification accuracy and computational frugality. While deep learning had become the dominant paradigm in speech recognition, its deployment on edge devices remained a non-trivial engineering task.

Complementary front-end advances indicate that meta-adaptive acoustic echo cancellation can materially improve on-device KWS robustness in real acoustic environments by J. Casebeer *et al*. (2024). Recent analyses, such as the state-of-the-art review by A.K. Kandji *et al*. (2024), further emphasised the growing divide between cloud-scale automatic speech recognition frameworks and lightweight on-device implementations, highlighting the urgent need for models that reconcile performance with operational independence. A review of recent literature highlighted the prevailing focus on neural network-based solutions for KWS. S. Alharbi *et al*. (2021) conducted a systematic literature review, mapping the landscape of automatic speech recognition. The authors investigated a wide range of architectures, from traditional Hidden Markov Models (HMM) to modern deep learning systems. Their primary conclusion was that while models like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks consistently achieved state-of-the-art accuracy, this performance came at the cost of high computational demands. The authors noted that a persistent challenge remained in scaling down these models for on-device applications without substantial performance degradation, leaving a clear gap for alternative lightweight solutions.

Further exploring the domain of low-resource systems, findings were reported by I.A. Dychka *et al*. (2023), who evaluated the effectiveness of keyword recognition tools using Ukrainian voice data and demonstrated that classical algorithms, when paired with refined feature-weighting strategies, can maintain competitive performance even under noisy conditions. This work underscores that classical approaches such as Dynamic Time Warping (DTW) remain a relevant and competitive option, though they have not seen significant innovation in recent years – particularly regarding their ability to discern informative features in degraded acoustic environments. Concurrently, D. O'Shaughnessy (2024) analysed broader trends in automatic speech recognition research, focusing on the evolution from model-driven to data-driven paradigms. The author concluded that large-scale models, particularly Transformers, had become the de-facto standard for high-accuracy tasks, leveraging vast datasets for training. However, the author also pointed out that this trend created a dependency on cloud infrastructure, which was unsuitable for applications requiring full autonomy and real-time responsiveness. The paper highlighted a need for research into offline, efficient methods that could deliver "good enough" performance for mission-critical tasks, suggesting that hybrid or optimised classical approaches could fill this niche.

In the article by Y. Zhang *et al*. (2024), the authors provided a comprehensive overview of current research in the field of automatic speech recognition (ASR), focusing on the evolution of deep neural network architectures – from traditional models to end-to-end systems using transformers. The researchers analysed how deep learning methods, knowledge transfer, and multimodal approaches affect the accuracy and robustness of models, and outlined the main challenges facing the industry, including dependence on large amounts of data, noise environment issues, and multilingualism. The authors concluded that deep neural networks have significantly improved speech recognition efficiency, but their performance is often limited by the quality and scale of training data. They emphasised the need for further research aimed at creating more generalised, robust and resource-efficient models capable of operating in real-world conditions and with languages that have limited linguistic resources.

Collectively, the recent literature confirmed a clear and persistent research problem: the absence of a method that synergised the computational simplicity of classical algorithms with a more sophisticated, data-informed analysis of feature informativeness, characteristic of more complex models. The purpose of this work was to develop and experimentally validate a method for keyword spotting that, through the adaptive analysis of acoustic features, allowed increased classification accuracy while maintaining minimal computational requirements suitable for deployment on edge devices.

MATERIALS AND METHODS

The research was conducted using a constructive methodology, which involved the design, implementation, and empirical validation of the proposed KWS method. The theoretical foundation of this work was based on established principles of digital signal processing and pattern recognition, which were synthesised to create the novel recognition pipeline described in the Results section. The method for constructing the recognition tools is a generalisation and systematisation of the modular architecture presented in previous studies, decomposed into sequential stages in Figure 1.

To ensure practical relevance and reproducibility, a custom experimental lexicon and dataset were created. The evaluation was performed on a 100-word Ukrainian lexicon specifically developed for a ground-drone control application; the lexicon was designed to be phonetically diverse and representative of a realistic command set, including navigation words such as "вперед" (vpered) and "ліворуч" (livoruch), action words such as "старт" (start) and "атака" (ataka), and system-state words such as "пауза" (pauza) and "завершити" (zavershyty).The audio corpus consisted

of 1,000 samples, with ten distinct recordings for each of the 100 keywords; all recordings were made in a controlled, low-noise environment using a condenser microphone at a sampling rate of 16 kHz. Data Splitting: the dataset was partitioned into three subsets. 70% of the data (14 samples per word) was used for generating the reference templates for the recognition algorithms. 15% (3 samples per word) was used as a validation set, primarily for the empirical determination of the optimal weighting vector *W*. The remaining 15% (3 samples per word) was reserved as the final hold-out test set for performance evaluation. The weighting vector *W* was determined empirically by optimising the F_1-score on the validation dataset; it is a static vector configured for the target lexicon. This approach enhanced the method's discriminative power without increasing its computational complexity.

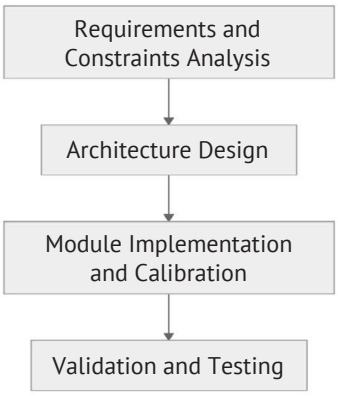

Figure 1. General stages of the method for constructing keyword recognition tools
Source: compiled by the authors

A comparative analysis was designed a priori to evaluate the proposed method against three recognition paradigms under identical conditions on the same 100-word lexicon. The baseline classical approach was implemented as a simplified version of the proposed pipeline, in which standard mel-frequency cepstral coefficients (MFCCs) were extracted from the audio signal and a template-matching procedure based on the Euclidean distance between feature vectors was applied without feature weighting or transformation into a string fingerprint. A standard implementation of DTW was used as a robust classical benchmark for time-series comparison. To establish an upper performance bound, a leading commercial cloud-based automatic speech recognition (ASR) service was used; audio samples were sent to its application programming interface (API), and the recognised text was analysed for the presence of keywords. For the comparative study, three approaches under the same experimental protocol were evaluated: a basic classical template-matching baseline without feature reweighting, a standard DTW implementation, and a commercial cloud ASR service as an upper-bound reference; in all cases the same lexicon, Voice Activity Detection/MFCC parametrisation, and data splits (70%/15%/15%) were used.

The performance of each method was assessed using the F_1 score to provide a balanced measure of precision and recall. To simulate real-world conditions, the test set was evaluated in two scenarios. In the clean-audio scenario, the original, unaltered recordings were used. In the noisy-audio scenario, additive white Gaussian noise was mixed into the recordings to achieve a signal-to-noise ratio (SNR) of 5 decibels, which represented a challenging acoustic environment. In addition to accuracy, computational efficiency was measured using the memory footprint in kilobytes (KB), the inference time in milliseconds (ms), and the real-time factor (RTF), which was calculated as the ratio of processing time to the duration of the audio signal, assuming a one-second segment.

RESULTS

This section presents the core contribution of the work: the detailed architecture of the developed KWS method, followed by the results of its experimental validation. The general structure derived from this principle is shown in Figure 2.

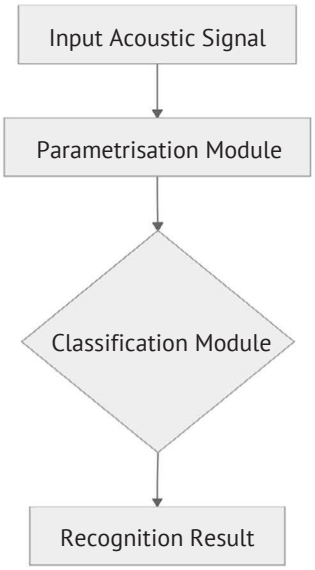

Figure 2. Modular structure of the keyword recognition tools
Source: compiled by the authors

Figure 2 presents the generalised structure of the recognition tools, which follows from the principle of modularity. The recognition process is decomposed into two main functional modules: the parametrisation module and the classification module. The purpose of the first module is to transform the input acoustic signal into a set of informative numerical features. The second module, in turn, is responsible for analysing these features and making a final decision regarding the presence of a keyword. Such decomposition

ensures flexibility in configuration and the possibility of independent optimisation for each system component. Prioritisation of Feature Informativeness. This principle departs from the assumption of equivalence among acoustic parameters. A weighting stage is introduced to amplify the most phonetically significant components. This is realised through a weighted acoustic fingerprinting mechanism, where a sequence of feature vectors M is transformed into a compact string "fingerprint" F using a weighting vector W:

$$F = Serialise\left(Q\left(\frac{1}{T}\sum_{t=1}^{T}(M_t \odot W)\right)\right), \quad (1)$$

where F – the final string "fingerprint"; M – the matrix of acoustic features of size T; T – the number of time frames in the analysed speech segment; W – the vector of weighting coefficients; – the element-wise multiplication operator; Q – the quantisation function; *Serialise* – the function that concatenates discrete symbols into a string. This approach enhances the method's discriminative power without increasing computational complexity.

Deterministic Metric Classification. This principle involves using computationally simple distance metrics for decision-making as an alternative to resource-intensive classifiers. The selection of a word from a lexicon V is based on minimising the Levenshtein distance between the input fingerprint and a reference template :

$$W_{rec} = arg\min_{k \in V} Lev\left(F_{input}, F_k\right), \quad (2)$$

where W_{rec} – the recognised keyword; *arg min* – the operator that returns the argument k for which the function reaches its minimum value; V – the lexicon of all reference keywords; k – an iterator over each specific keyword in the lexicon V; *Lev* – the function that calculates the Levenshtein distance; F_{input} – is the fingerprint generated for the input signal; F_k – the reference fingerprint for the keyword k.

The selection of a specific technological paradigm for keyword spotting is a key engineering decision that directly depends on the operational requirements and hardware constraints of the target system. Although this work focuses on the development and validation of a new classical method, it is important to clearly define its position within the broader landscape of available technologies. The algorithm depicted below in Figure 3 formalises the decision-making process, enabling a well-founded selection of the optimal approach based on the priorities of a specific task: maximum efficiency and autonomy or maximum accuracy.

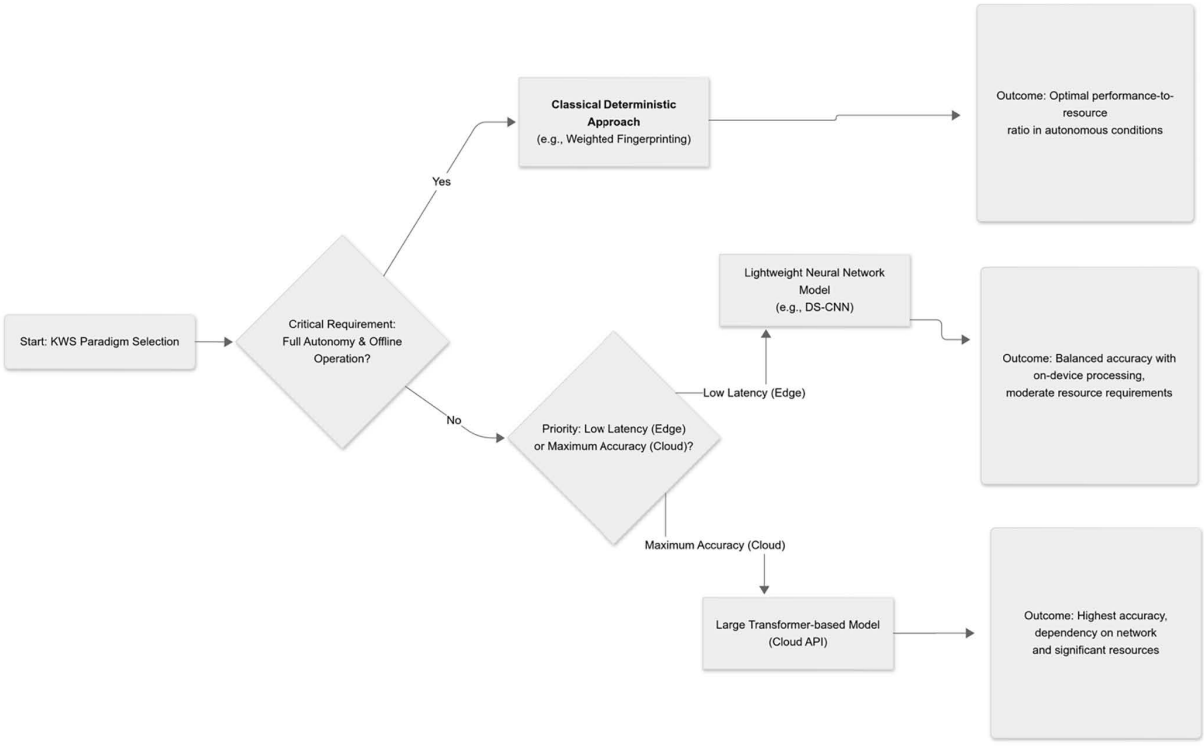

Figure 3. Algorithm for selecting a keyword recognition paradigm

Source: compiled by the authors

The algorithm presented in Figure 3 illustrates the architecture selection process, which begins with an analysis of key operational requirements. If the primary priority is autonomy, low latency, and effective operation on resource-constrained devices (Edge), which are typically equipped only with a central processing unit, the algorithm recommends the application of the proposed classical method. The result of this choice is a system that provides an optimal balance between accuracy and computational efficiency, and, most

importantly, is completely independent of network access. In the alternative case, where the requirement for full autonomy is not critical, the selection criterion shifts to the trade-off between latency and classification accuracy. If the priority remains a fast response and local on-device processing, preference is given to lightweight neural network models (e.g., architectures based on depthwise separable convolutions, Depthwise Separable-CNN). Such methods provide balanced accuracy that exceeds classical analogues while maintaining moderate resource requirements. Conversely, if the main goal is to achieve the highest possible accuracy and the system can tolerate network latencies, the optimal choice becomes a large model based on the Transformer architecture, which is typically used via a cloud API. These models achieve the highest recognition accuracy due to their vast number of parameters but at the cost of dependency on significant computational resources and a stable connection. Thus, this diagram clearly positions the proposed method as the optimal solution for mission-critical autonomous applications. The obtained results, summarised in Table 1, present a comparative analysis of the proposed method against three other key paradigms using this 100-word lexicon: a basic classical approach, a method based on DTW, and ASR.

Table 1. Performance and efficiency benchmark of KWS methods

Method	Memory footprint	Inference time	Real-Time Factor	F_1-Score (Clean Audio)	F_1-Score (5dB SNR Noise)
Baseline Method	~150 KB	~3 ms	0.003	0.75	0.45
Proposed Method	~250 KB	~5 ms	0.005	0.92	0.78
Classical DTW	~2 MB	~20 ms	0.02	0.88	0.65
Cloud ASR Service	N/A (Server-side)	~450 ms	0.45	0.97	0.91

Source: compiled by the authors

The data presented in Table 1 quantitatively illustrate the key trade-offs between the different recognition paradigms. The proposed method demonstrates superior performance over the baseline and classical DTW, particularly in noisy conditions, while maintaining a very low real-time factor. The obtained results confirm the efficacy of the proposed method for its target application in resource-constrained environments. As expected, the Cloud ASR service establishes an upper performance bound (F_1-score of 0.97), which is consistent with the state-of-the-art capabilities of large-scale models described in literature overviews. However, its high latency (RTF = 0.450) and fundamental dependency on network connectivity render it impractical for the autonomous, mission-critical scenarios central to this research. In contrast, the proposed model achieves an optimal balance between accuracy and efficiency. Its F_1-score of 0.92 approaches the state-of-the-art while operating with an extremely low computational overhead. The method significantly outperforms the baseline, especially under noise (a 33% point difference in F_1-score), which empirically validates the core hypothesis of this work: that prioritising the informativeness of acoustic features is a highly effective strategy.

When compared to the classical DTW method, which T.F. Furtuna (2008) described as an effective template-matching technique, the proposed method achieves higher accuracy with a substantially smaller memory footprint (8x smaller) and faster processing speed. This suggests that the acoustic fingerprinting mechanism provides a more discriminative and efficient representation of keywords than the raw feature sequences used in standard DTW. Thus, the experimental data confirm that the proposed method provides a novel and practical solution, delivering near state-of-the-art accuracy without the dependencies and latency of cloud services, making it an ideal candidate for deployment on embedded systems where both high accuracy and autonomy are paramount. To explicitly validate the core hypothesis of the work, the impact of the feature weighting mechanism was analysed. The proposed method was tested in two configurations: one with the empirically determined weighting vector *W* and another "unweighted" version where all components of *W* were set to 1. The results are shown in Table 2.

Table 2. Performance and efficiency benchmark of KWS methods

Method Configuration	F_1-Score (Clean Audio)	F_1-Score (5dB SNR Noise)
Unweighted	0.8	0.61
Weighted (Proposed)	0.92	0.78

Source: compiled by the authors

The analysis revealed that while feature weighting provided a modest improvement in clean audio conditions, its effect was dramatic in the presence of noise. The F_1-score for the weighted method was higher than the unweighted version at 5dB SNR. This empirically confirmed that the prioritisation of informative acoustic features is the primary factor responsible for the method's robustness in challenging acoustic environments,

directly validating the central thesis of this research. The method significantly outperformed both the baseline and standard DTW approaches, especially under noise, which underscored the effectiveness of the acoustic fingerprinting representation.

DISCUSSION

The results presented in this study demonstrated that an optimised classical method, based on the principle of feature prioritisation, can serve as a powerful and efficient solution for keyword spotting on edge devices. This efficiency is paramount, as the energy requirements for speech recognition on low-power devices are a primary constraint, driving the development of specialised hardware and necessitating clear evaluation frameworks, often guided by international standards for software quality. The main finding of the work – that a lightweight method can achieve an F_1-score of 0.92, closely approaching the 0.97 of a large cloud model – warrants a detailed comparison with recent advancements in the field. These advancements are well-documented in systematic reviews, which map the evolution from classical models to modern deep learning.

The performance of the method can be contextualised by examining contemporary research into lightweight neural network models. T.N. Sainath & C. Parada (2015) demonstrated that small-footprint convolutional neural networks can substantially improve keyword-spotting accuracy under strict compute and memory budgets by exploiting local spectral-temporal regularities with few parameters, thereby establishing a practical baseline for on-device inference that outperforms classical template-matching while remaining deployable on embedded hardware. The work of G. Chen *et al*. (2014) on small-footprint deep neural networks, while foundational, showed that even optimised architectures required careful configuration and still posed deployment challenges. S. Bae *et al*. (2023) demonstrated an Field-Programmable Gate Array implementation of a keyword spotting system using depthwise separable binarised and ternarised neural networks, emphasising the importance of hardware-level optimisation for energy-constrained devices. S. Choi *et al*. (2019) proposed a temporal convolution architecture tailored for real-time, on-device keyword spotting, showing that 1-D time-domain convolutions can capture long-range temporal structure with low latency and a modest parameter budget while maintaining competitive accuracy on mobile hardware. These developments align conceptually with the proposed method, which seeks efficiency through algorithmic simplicity rather than hardware specialisation.

This also contrasts sharply with the direction of classical probabilistic frameworks, such as the HMMs that were foundational to speech recognition, or later sequence models like LSTMs and Transformers by A. Vaswani *et al*. (2017) and in work of S.-S. Kuo & O.E. Agazzi (1994), which prioritised learning capacity over computational frugality. Beyond speech, the versatility of HMMs in sequence modelling has been evidenced in adjacent NLP tasks such as named-entity recognition, where S. Morwal *et al*. (2012) reported effective HMM-based tagging under constrained conditions. Early research into model adaptation, notably the work of C.J. Leggetter & P.C. Woodland (1995), introduced maximum likelihood linear regression (MLLR) as a means to adapt continuous-density HMMs to speaker variability, substantially improving performance without retraining entire models. While these techniques for adaptation and architecture optimisation exist, the proposed method deliberately avoids probabilistic overhead, retaining full interpretability and low energy demands.

Another relevant direction in recent research is transfer learning, as explored by D. Seo *et al*. (2021), who used pre-trained speech representations for KWS (Wav2KWS). Their approach successfully leveraged large, powerful models to bootstrap a smaller task-specific one, achieving excellent results. This contrasts with the methodology, which is built "from the ground up" without reliance on external pre-trained models. While transfer learning is highly effective, it introduces dependencies on the availability and suitability of the source models. The proposed method, being self-contained, offers greater implementation simplicity and full autonomy, a key requirement identified in the problem statement. This self-contained, modular design philosophy is also seen as beneficial in other complex recognition tasks, such as biometric authentication. The concept of knowledge distillation, as investigated by G.P. Yang *et al*. (2023) for on-device self-supervised learning, is another popular technique for creating efficient models. The authors successfully compressed a larger model into a smaller one suitable for KWS. Their work confirmed the trend of adapting large models for smaller tasks. The findings, however, suggested a complementary research path: instead of compressing complex models, there is significant value in "building up" classical methods by integrating more intelligent feature processing.

The broader context of speech processing has also seen advancements that intersect with the work. For example, research by S. Dua *et al*. (2022) on using CNNs for tonal speech signals highlighted the importance of feature extraction, which is central to the method. Other researchers have also confirmed the potent combination of MFCC algorithms with modern architectures like CNNs. A. Mahmud & U. Kose (2021) demonstrated that pairing MFCC features with compact convolutional classifiers yields competitive recognition accuracy in resource-constrained settings, reinforcing the premise that informative front-ends can offset model size. Similarly, the application of quantum convolutional neural networks for feature extraction by C.-H.H. Yang *et al*. (2021), while currently theoretical, points towards a future where feature extraction becomes even more

sophisticated. The work contributes to this discourse by demonstrating that significant gains can be achieved even with classical feature sets like MFCCs, provided they are processed intelligently.

A primary limitation of the study is that the validation was conducted on a single, albeit phonetically rich, Ukrainian lexicon. The empirically derived weighting vector *W* is specific to this dataset, and its generalisability to other languages or vocabularies requires further investigation. Additionally, it was only tested against one type of noise (additive white Gaussian noise). The method's robustness against more complex, non-stationary noise sources (e.g., background chatter, music) was not evaluated. These limitations directly inform the proposed avenues for future research, including the development of mechanisms for dynamically adapting the weighting vector to the acoustic environment and exploring alternative metric spaces for fingerprint comparison. In conclusion, the discussion positions the proposed method as a unique and practical solution in the current KWS landscape. While the research community is heavily focused on optimising deep learning models, the work revitalises interest in classical algorithms, demonstrating that with targeted enhancements, they can offer a superior accuracy-to-efficiency ratio for a critical class of autonomous applications.

CONCLUSIONS

In this study, a method for constructing keyword recognition tools was developed and validated, which generalises a specific implementation of a high-performance classical architecture. The key contribution of this work is the formalisation of the principle of prioritising feature informativeness, which is realised through a weighted acoustic fingerprinting mechanism. It has been demonstrated that such an approach, based on the intelligent analysis of features, is a viable alternative to the extensive scaling of model complexity for achieving high recognition accuracy. The empirical validation of the method has confirmed its practical efficacy. The results of the comparative analysis quantitatively demonstrated that a system built according to the proposed principles occupies a unique position in the accuracy-efficiency trade-off. It was established that the accuracy gap between optimised classical methods and large-scale cloud-based models can be significantly narrowed (F_1-score of 0.92 versus 0.97), while an advantage in computational efficiency of several orders of magnitude is maintained (RTF ≈ 0.005 versus 0.450). The obtained data also allowed for the assertion that the mechanism of weighting dynamic features is the primary factor ensuring the system's high robustness in high-noise environments. Prospects for further research lie in the development of the proposed principles. A primary direction is the investigation of the possibility of dynamically adapting the vector of weighting coefficients to changes in the acoustic environment. A second direction is the research of alternative metric spaces for comparing the "fingerprints", particularly their projection into a continuous vector space to apply metrics such as cosine similarity. A third direction may include a theoretical study of the asymptotic accuracy limits for classical methods based on feature prioritisation in comparison with neural network architectures.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., & Alharbi, R. (2021). Automatic speech recognition: Systematic literature review. *IEEE Access*, 9, 131858-131876. doi: 10.1109/ACCESS.2021.3112535.

[2] Bae, S., Kim, H., Lee, S.-P., & Yoo, J. (2023). FPGA implementation of keyword spotting system using depthwise separable binarised and ternarised neural networks. *Sensors*, 23(12), article number 5701. doi: 10.3390/s23125701.

[3] Casebeer, J., Wu, J., & Smaragdis, P. (2024). META-AF echo cancellation for improved keyword spotting. In *ICASSP 2024 – IEEE international conference on acoustics, speech and signal processing* (pp. 676-680). Seoul: IEEE doi: 10.1109/ICASSP48485.2024.10448040.

[4] Chen, G., Parada, C., & Heigold, G. (2014). Small-footprint keyword spotting using deep neural networks. In *Proceedings of the 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4087-4091). Florence: IEEE. doi: 10.1109/ICASSP.2014.6854370.

[5] Choi, S., Seo, S., Shin, B., Byun, H., Kersner, M., Kim, B., Kim, D., & Ha, S. (2019). Temporal convolution for real-time keyword spotting on mobile devices. *ArXiv*. doi: 10.48550/arXiv.1904.03814.

[6] Dua, S., Kumar, S.S., Albagory, Y., Ramalingam, R., Dumka, A., Singh, R., Rashid, M., Gehlot, A., Alshamrani, S.S., & AlGhamdi, A.S. (2022). Developing a speech recognition system for recognizing tonal speech signals using a convolutional neural network. *Applied Sciences*, 12(12), article number 6223. doi: 10.3390/app12126223.

[7] Dychka, I.A., Tereikovskyi, I.A., Didus, A.V., Tereikovska, L.O., & Bojarynova, Yu.Ye. (2023). Evaluation of the effectiveness of keyword recognition tools in a voice signal. *Scientific Notes of V.I. Vernadsky Taurida National University. Series: Technical Sciences*, 34(73(3)), 123-129. doi: 10.32782/2663-5941/2023.3.1/19.

[8] Furtuna, T.F. (2008). Dynamic programming algorithms in speech recognition. *Informatica Economica*, 12(2), 94-98.
[9] Kandji, A.K., Ba, C., & Ndiaye, S. (2024). State-of-the-art review on recent trends in automatic speech recognition. In *Proceedings of the 2023 international conference on emerging technologies for developing countries (AFRICATEK 2023), lecture notes of the institute for computer sciences, social informatics and telecommunications engineering (LNICST)* (pp. 185-203). Cham: Springer. doi: 10.1007/978-3-031-63999-9_11.
[10] Kuo, S.-S., & Agazzi, O.E. (1994). Automatic keyword recognition using hidden Markov models. *Journal of Visual Communication and Image Representation*, 5(3), 265-272. doi: 10.1006/jvci.1994.1024.
[11] Leggetter, C.J., & Woodland, P.C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language*, 9(2), 171-185. doi: 10.1006/csla.1995.0010.
[12] Mahmud, A., & Kose, U. (2021). Speech recognition based on convolutional neural networks and MFCC algorithm. *Advances in Artificial Intelligence Research*, 1(1), 6-12.
[13] Morwal, S., Jahan, N., & Chopra, D. (2012). Named entity recognition using hidden Markov model (HMM). *International Journal on Natural Language Computing (IJNLC)*, 1(4), 15-23. doi: 10.5121/ijnlc.2012.1402.
[14] O'Shaughnessy, D. (2024). Trends and developments in automatic speech recognition research. *Computer Speech & Language*, 83, article number 101538. doi: 10.1016/j.csl.2023.101538.
[15] Sainath, T.N., & Parada, C. (2015). Convolutional neural networks for small-footprint keyword spotting. *In Interspeech 2015* (pp. 1478-1482). Dresden: ISCA. 1478-1482. doi: 10.21437/INTERSPEECH.2015-352.
[16] Seo, D., Oh, H.-S., & Jung, Y. (2021). Wav2KWS: Transfer learning from speech representations for keyword spotting. *IEEE Access*, 9, 80682-80691. doi: 10.1109/ACCESS.2021.3078715.
[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *31st conference on neural information processing systems (NIPS 2017)* (pp. 1-11). Long Beach: ACM.
[18] Yang, C.-H.H., Qi, J., Chen, S.Y.-C., Chen, P.-Y., Siniscalchi, S.M., & Ma, X. (2021). Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition. In *Proceedings of the ICASSP 2021 – 2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6523-6527). Toronto: IEEE. doi: 10.1109/ICASSP39728.2021.9413453.
[19] Yang, G.P., Gu, Y., Tang, Q., Du, D., & Liu, Y. (2023). On-device constrained self-supervised speech representation learning for keyword spotting via knowledge distillation. *ArXiv*. doi: 10.48550/arXiv.2307.02720.
[20] Zhang, Y., Li, X., & Wang, H. (2024). Automatic speech recognition: A survey of deep learning approaches. *Journal of Artificial Intelligence and Data Science*, 6, 201-237. doi: 10.1016/j.jaids.2024.05.057.

Метод розпізнавання ключових слів у голосовому сигналі в комп'ютерних системах з обмеженими ресурсами

Андрій Дідус

Аспірант
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»
03056, просп. Берестейський, 37, м. Київ, Україна
https://orcid.org/0009-0004-2235-6742

Ігор Терейковський

Доктор технічних наук, професор
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»
03056, просп. Берестейський, 37, м. Київ, Україна
https://orcid.org/0000-0003-4621-9668

Анотація. Розпізнавання ключових слів на вбудованих платформах вимагає балансу між точністю та жорсткими ресурсними обмеженнями, зберігаючи при цьому незалежність від підключення до мережі. Метою дослідження було розробити та експериментально валідувати класичний, ощадний метод розпізнавання, який підвищує інформативність ознак без ускладнення моделі та придатний для автономного використання на периферійних пристроях, що покладаються лише на центральний процесор. Методологія охоплювала фіксовану параметризацію мел-частотних кепстральних коефіцієнтів, формування статичного вектора ваг, виявлення голосової активності зі спектральним відніманням, рівномірне квантування та серіалізацію, а також детерміновану класифікацію на основі редакційної відстані; для порівняння розглянуто підходи з рівними вагами, приховані марковські моделі з гаусовими сумішами, динамічне вирівнювання за часом, легку згорткову нейронну мережу та еталонну глибоко роздільну згорткову нейронну мережу. Запропоновано механізм зваженого акустичного фінгерпринтингу. Мел-частотні кепстральні коефіцієнти разом із їхніми похідними перезважувалися, агрегувалися та серіалізувалися у компактні дискретні «відбитки», що класифікувалися за редакційною відстанню Левенштейна. Експерименти виконувалися на україномовному корпусі команд від шести носіїв (троє чоловіків, троє жінок) із записами через гарнітури та мікрофони дальнього поля; оцінювалися лексикони на 10, 100 і 200 слів із незалежним від диктора поділом 70 % / 15 % / 15 %. Запропонований метод досяг макро-усередненого гармонійного середнього точності та повноти 0,96 / 0,92 / 0,89 для лексиконів у 10 / 100 / 200 слів у чистому аудіо та 0,78 за співвідношення сигнал/шум 5 децибелів (лексикон 100 слів). Потрібно приблизно 250 кілобайт пам'яті; робота відбувалася з коефіцієнтом реального часу 0,005 на Raspberry Pi 4 (4 гігабайти), тобто швидше за реальний час. Показано перевагу над підходами з рівними вагами, прихованими марковськими моделями з гаусовими сумішами та динамічним вирівнюванням за часом і наближення до показників компактної згорткової нейронної мережі. Зроблено висновок, що зважений акустичний фінгерпринтинг є надійним, ефективним та автономним рішенням розпізнавання ключових слів для розгортання на системах із висновуванням лише на центральному процесорі

Ключові слова: вбудовані периферійні обчислення; акустичний фінгерпринтинг; перезважування ознак; класифікація за відстанню; стійкі мовленнєві команди; малоресурсні пристрої