Customized Wake Word Integration in ANC-Enabled Headphones: Enhancing Assistive Technology for Noise-Sensitive Employees in Office

Jan Schmalfuß-Schwarz
jan.schmalfuss-schwarz@tu-dresden.de
TUD Dresden University of Technology
Germany

David Gollasch
david.gollasch@tu-dresden.de
TUD Dresden University of Technology
Germany

Meinhardt Branig
meinhardt.branig@tu-dresden.de
TUD Dresden University of Technology
Germany

Christin Engel
christin.engel@tu-dresden.de
TUD Dresden University of Technology
Germany

Gerhard Weber
gerhard.weber@tu-dresden.de
TUD Dresden University of Technology
Germany

ABSTRACT

This paper explores the personalization of wake words in Active Noise Cancelling (ANC) enabled headphones designed as assistive technology for noise-sensitive individuals. We focus on the incorporation and assessment of intended wake words, categorizing them based on fillers and first names. Given the language-dependent nature of this approach, our research primarily addresses the English language. We propose an evaluation mechanism to determine the efficacy of potential wake words, predicting their recognition accuracy and length. The study involves a comparative analysis between combinations of fillers and first names, aiming to identify the optimal pairing. We offer recommendations on the best-performing combinations, enhancing the reliability and user experience of ANC headphones as a supportive tool for noise-sensitive individuals. Our findings aim to provide a robust framework for developing highly responsive and personalized wake word systems, tailored to the unique needs of noise-sensitive individuals in office environments.

KEYWORDS

Personalised Wake Words, Assistive Technology, Noise Sensitivity, Smart Headphones, Active Noise Cancelling, Office Environment

ACM Reference Format:
Jan Schmalfuß-Schwarz, David Gollasch, Meinhardt Branig, Christin Engel, and Gerhard Weber. 2024. Customized Wake Word Integration in ANC-Enabled Headphones: Enhancing Assistive Technology for Noise-Sensitive Employees in Office. In *Proceedings of Mensch und Computer 2024 (MuC '24), September 01–04, 2024, Karlsruhe, Germany*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3670653.3677489

MuC '24, September 01–04, 2024, Karlsruhe, Germany
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0998-2/24/09
https://doi.org/10.1145/3670653.3677489

1 INTRODUCTION

Focused work in the office is increasing for work activity in our society. Depending on the person's sensitivity to noise, this requires a quiet workplace or assistive technology to shield against acoustic stimuli. A potential target group could be autistic people, as they are sensitive to auditory stimuli. Radun et al. have demonstrated that Active Noise Canceling (ANC) and masking noises using headphones can be effective for noise-sensitive people to improve concentration at work [6]. They also found that, for example, conversation from other people influences concentration negatively [7]. At the same time, EN ISO 10075-2:2000 describes that social interaction in the workplace is an important aspect to prevent monotony. Therefore, this paper aims to discuss the possibility of personalized wake-up phrases and to create an initial approach to evaluating potential wake-up phrases.

2 STATE OF THE ART

One of the earliest approaches in wake word recognition involves the use of LVCSR systems, which were researched as early as 1993 [13] and are characterized by their flexibility [10]. However, they require significant resources and can generate latencies [11]. In contrast, most digital assistants such as Apple's Siri or Amazon's Alexa use machine learning for wake word recognition. Examples of this include the use of convolutional-recurrent-attention networks [4] or the use of metadata in the recognition process [5], which are optimizations of already existing systems. An alternative technology option is to use HMMs (in conjunction with GMMs [8] or Deep Neural Networks (DNNs) [2]). However, these approaches use the Viterbi algorithm [12] for decoding, which can also be computationally intensive - depending on the structure of the HMMs [1]. In addition to the solution approaches described, deep learning can be used to recognize wake words. This approach is well suited for devices with memory and processing limitations [3].

Zhang et al. also discuss the necessity for the keyword spotting process to work on microcontrollers and have developed an approach using various neural networks. Additionally, the authors have published the code under an open-source license on GitHub so that it can be reused[1] and use a freely accessible data set for their implementation [2]. [14]

In contrast to the emphasis on methods for recognizing wake words, the evaluation of appropriate quality for personalized wake words is typically not the primary focus. An indication is provided by an interview published in 2021 with Darin Clark, Director of Business Development at SoundHound (a voice and conversation recognition technology company), and Kane Simms from VUX World (a voice user experience platform and community) on the topics of wake word recognition and the criteria for selecting wake words.[3] Within the interview Clark, for example, recommends choosing wake words that are four to five syllables long to make them easier to distinguish from other words. Therefore, short names require filler words, which, according to Clark, also means that only a direct response from an assistive system is recognized. As a counterexample, Clark cites mentioning the name of a company in a direct conversation, which can activate the system without a corresponding filler word. Additionally, Clark recommends using words with a large phonetic variation, where consonants and vowels alternate as much as possible.

3 CUSTOMIZED WAKE-UP PHRASES AND SCORING MECHANISM

In the following section, requirements are derived. This is done from the literature sources, with the office context being the focus of the work [9]. In addition, a suggestion for scoring personalized wake-up phrases is being developed. In this context, we will continue to discuss wake-up phrases, which consist of a filler word (in our case a greeting such as "Hello") and the person's first name.

3.1 Requirements

The use of headphones in the office aims to enable focused work. This can be supported by ANC or masking. At the same time, communication capability is a central aspect of work. The nature of communication varies greatly and depends on the specific context. Consequently, the requirements are diverse and sometimes conflicting.

Noise-Sensitive User. For focused work, headphones must prevent false wake word recognitions to avoid disruptions, while ensuring system responsiveness to maintain user acceptance and prevent isolation. The system should be portable, allowing users to move around the office without exposing themselves to increased stimulation. This portability requirement also impacts system performance, as additional large components should be avoided.

Office Environment. The system must consider the needs of both the headphone user and surrounding colleagues. It should adapt to different forms of greeting appropriate to company roles, hierarchies, and other factors. The system should work accurately and provide feedback when a corresponding wake-up phrase has been recognized. Wake-up phrases should be easy to learn or cover a broad range to decrease cognitive effort to start interaction, while avoiding the use of names alone as wake words to prevent unintended activations.

The requirements can be summarized into variance, performance, and accuracy. The characteristics of the requirements are context-sensitive and therefore dependent on the respective work activity and the working environment. However, these requirements often conflict, e.g. as high variance combined with high accuracy can lead to reduced system performance, limiting possible applications. Similarly, focusing on performance and accuracy may limit variance options, and focusing on performance and variance would reduce accuracy. Therefore, evaluating potential wake-up phrases is crucial to achieve optimal overall system performance. Given the high effort required for data acquisition and the need for easy implementation and expandability, a prior assessment of suitable wake-up phrases in the specific context of the working environment is necessary.

3.2 Rating of Wake-up Phrases based on Filler and First Name

Based on the requirements, we developed a score (cf. Fig. 1) that uses Eudex[4] as a phonetic hashing algorithm and compares filler words and names. Higher scores indicate greater word variance. Based on this, we use the score and divide it by the number of syllables of the wake-up phrase. Moreover, wake-up phrases with a size of at least 6 syllables receive an additional factor of 1.5, and phrases with more than 8 syllables receive a factor of 2. This approach is based on the assumption that wake-up phrases should not exceed a certain length. The calculated score for the wake-up phrases examined in the paper can be found in Table 1.

4 EVALUATION OF THE RATING

To evaluate wake-up phrases, various English greetings with the names Charlie and Olivia were combined. The two names were chosen based on their different number of syllables (Char-lie / O-li-vi-a) and their similar overall length. The chosen greetings are "Hi" with one syllable, "Hello" with two syllables, "Howdy" with two syllables, "Hiya" with two syllables, and "Nice to see you" with four words, each consisting of one syllable. The two chosen names were permuted with all greetings, creating a set of 10 different wake-up phrases to prove the assumptions made (cf. Table 1).

4.1 Data Collection

To capture the audio data, a Python program for home use was written which, after a brief textual explanation as console output, asked the participants first to provide demographic data such as age, gender, native language, and microphone used by entering it into the console. For the native language, German and English were offered as possible options, and, concerning the microphone used, the choice was given between internal and external microphones, a headset, and other microphones. This is followed by a microphone test in which the participants should say the greeting "Hi Charlie"

[1] GitHub "Keyword spotting on microcontrollers: https://github.com/ARM-software/ML-KWS-for-MCU [Last access: June 11, 2024]

[2] Speech command dataset by Google: https://research.google/blog/launching-the-speech-commands-dataset/ [Last access: June 11, 2024]

[3] N.N. CEO Keyvan Mohajer: "What You Need to Know About Wake Word Detection" (as of May 30, 2023) At: https://www.soundhound.com/voice-ai-blog/what-you-need-to-know-about-wake-word-detection/ [Last access: June 11, 2024]

[4] GitHub Eudex: https://github.com/remiadon/eudex [Last access: June, 11, 2024]

$$score = \frac{eudex}{syllables * c_{\text{syll}}} \quad \text{with} \quad c_{\text{syll}} = \begin{cases} 1, & \text{for } 3 \leq \text{syllables} < 6 \\ 1.5, & \text{for } 6 \leq \text{syllables} < 8 \\ 2, & \text{else} \end{cases} \quad , \quad \forall \, \text{syllables} \geq 3$$

Figure 1: Function to determine the score of a wake-up phrase based on the phonetic comparison of the filler and the name using Eudex and the number of syllables.

and confirm that the microphone is working. The 10 different wake-up phrases were then recorded, with each participant having to say each phrase 10 times. The start of the individual recordings was initiated by pressing Enter, which resulted in a 2-second sound recording. The participants were assured that the audio recordings would be processed in accordance with data protection regulations, which precluded publication of the recordings. In our opinion, this is relevant because we received up to 200 seconds of audio from each participant, which could otherwise be used for other purposes beyond detecting wake-up phrases.

4.2 Participants

A total of 13 people took part in data acquisition. 8 of the 13 people were male and 5 were female. The native language of all participants is German, and the age is between 22 and 62 years. For the recordings, internal laptop microphones were used 12 times, and an external microphone once. When selecting the participants, we also made sure that everyone had a basic level of English.

4.3 Data Preprocessing

In the first preprocessing step, all collected audio files were listened to and examined for abnormalities such as incomplete recordings due to missing beginnings or ends of the wake-up phrase, empty audio files due to failed recordings, hang-ups in the participants' pronunciation, and masking noises during the wake-up phrase. These audio files were sorted out. Therefore, three data sets were eliminated because, once in general had a large number of recordings with errors, and twice, had a large amount of recording errors in relation to a specific phrase. All other data sets were continued to be used. Furthermore, in cases where we need to remove a recorded wake-up phrase from one participant, we also eliminate one wake-up phrase from all other sets associated with that participant to maintain uniform distribution. This resulted in a data set with a total of 910 audio samples of 10 different people, which were evenly distributed across the 10 wake-up phrases. The number of audio files per participant fluctuates between 100 and 60 recordings. After manual preprocessing, further machine processing was done using a Python script, where the audio files were normalized and the length was cut based on a volume threshold. Finally, a manual run was carried out again, in which we listened to the processed audio files. During this step, we re-edited audio files, for example, if they were not cut correctly. This led to a manual revision of 56 audio files.

4.4 Training of the Model

For the implementation of the model, we used the by Zhang et al. in Python developed training script[5]. Based on this, a DNN architecture was chosen, which consists of 3 fully-connected layers, each with 128 neurons. The training itself included 18,000 steps. In the first iteration, additional audio files from Google's speech_command data[6] set were integrated. Therefore, we use the words "Bed", "Bird", "Cat", "Dog", "Happy", "House", "Marvin", "Sheila", "Tree", and "Wow" and reduce the datasets of each to 45 samples to avoid an unbalanced data set. To validate the individual wake-up phrases, we use 14 different data sets for each phrase, as well as 14 for all unlabeled words. In a further step, we trained single models purely on the self-collected data, with each word being used as a labeled word and all others being considered unlabeled. Therefore, we used 18 data sets to validate the individual wake-up phrases. After a first run, we reduced the entire data set and removed the wake-up phrase "Hiya", as it influenced the results due to different pronunciations: "Heya" as well as "Hiya", and went through both training sections again.

4.5 Results

One challenge in the current study is the amount of data. Based on the small data set, the accuracy, recall, and precision of recognizing the individual wake-up phrases is still imprecise and is influenced by other effects. Moreover, the greeting "Hiya" was removed inside a second run of both trainings, which raised the mean of the F1-score from 0.75 to 0.93 for the first training and the mean of the second training accuracy from 0.92 to 0.93. In addition, the small data set also limits the potential size of the test data set, which means that a smaller-scale determination of the recall, precision, and accuracy is not possible.

The confusion matrix of the first iteration showed that the filler words are highly relevant. For example, all wake-up phrases relating to "Nice to see you Olivia" were recognized correctly, but two test files from "Nice to see you Charlie" were also assigned to the phrase "Nice to see you Olivia". In no other case was a phrase incorrectly assigned to another phrase twice. The F1 score derived from the equally weighted precision and recall fluctuates between 0.929 and 0.966. For "Charlie", the filler "Hi," "Hello," and "Howdy" achieved the best values. In the case of "Olivia," the filler "Hi" and "Hello" achieved the best score. (cf. Table 1)

The current accuracy based on the second turn is as follows: "Hello Charlie", "Hello Olivia" and "Howdy Olivia" with 88.9 %; "Nice to see you Charlie", "Hi Olivia" and "Nice to see you Olivia" with 94.4 %

[5] GitHub "Keyword spotting on microcontrollers: https://github.com/ARM-software/ML-KWS-for-MCU [Last access: June 11, 2024]

[6] Speech command dataset by Google: https://research.google/blog/launching-the-speech-commands-dataset/ [Last access: June 11, 2024]

Table 1: Overview of the different Wake-up Phrases, their accuracy, recall, precision, and F1-Score in the first training as well as the accuracy of the second training, their count of syllables, their Eudex-Score and their aggregated score based on our approach

Wake-up Phrase	Recall 1. Training	Precision 1. Training	F1-Score 1. Training	Accuracy 2. Training	Syllables	Eudex- Score	Calculated Score
Hi Charlie	1.0	0.875	**0.933**	**1.0**	3	6	**2.0**
Hello Charlie	0.857	1.0	0.923	0.889	4	4	1.0
Howdy Charlie	0.857	1.0	0.923	**1.0**	4	7	**1.8**
Nice to see you Charlie	0.857	1.0	0.923	0.944	6	15	1.7
Hi Olivia	1.0	0.933	**0.966**	**0.944**	5	11	**2.2**
Hello Olivia	1.0	0,933	**0.966**	0.889	6	13	**1.4**
Howdy Olivia	1.0	0.929	0.929	0.889	6	12	1.3
Nice to see you Olivia	1.0	0.875	0.933	**0.944**	8	18	1.1

and "Hi Charlie" and "Howdy Charlie" with 100 %. (cf. Table 1)
In comparison, the Eudex score delivers values between 4 and 18, which ultimately leads to an overall score between 1.0 and 2.2 according to our defined approach. In this regard, "Hi Charlie" with 2.0 and "Howdy Charlie with 1.8 for Charlie and "Hi Olivia" with 2.2 and "Hello Olivia" with 1.4 for Olivia were the best scored wake-up phrases. (cf. Table 1)
Considering an existing correlation between the Eudex score, the number of syllables, and the accuracy or F1-score does not make sense at this early state.

4.6 Discussion

Based on the results, it should be noted that it is not possible to make concrete statements about the quality of the wake-up phrase score function. It seems obvious that the function provides a good guideline for suitable wake-up phrases, but the acquisition of a larger data set is necessary to validate and improve it.
Nevertheless, the training allows conclusions to be drawn about the quality. In the first iteration, the wake-up phrases "Hi Charlie" (14 TP and two times FP: "Hello Charlie" and unlabeled words) generated the best results for Charlie. The wake-up phrases "Hello", "Howdy", and "Nice to see you" achieve the same F1 score (in each case 12 TP and two times FN). Regarding Olivia, these are "Hi Olivia" (14 TP and one time FP: "Howdy Olivia") as well as "Hello Olivia" (14 TP and one time FP: unlabeled words). For "Charlie" and "Olivia", the highest calculated score achieves the best results for the F1-score of the first and the accuracy of the second training step. (cf. Table 1)
Moreover, initial assumptions can be made regarding the length of the wake-up phrase based on the recorded audio files. At the beginning of preprocessing, a total of 11 audio files were sorted out, which contain hang-up pronunciations. 7 of these refer to wake-up phrases with the filler "Nice to see you" and only 4 to all other wake-up phrases, with 3 of them being the rather unusual and often mispronounced "Hiya". Likewise, the 7 hang-up pronunciations are distributed between "Nice to see you Charlie" 4 times and "Nice to see you Olivia" 3 times and were made by 6 different participants. This suggests that long fill words should be avoided and that their length must therefore be taken into account by the score, as we already consider by dividing by the number of syllables and a coefficient.

5 CONCLUSION

Radun et al. showed in their study that headphones with ANC and/or for masking noise can have advantages for noise-sensitive people in the workplace [6]. At the same time, it is important to avoid isolation from the social environment. Therefore, it is necessary to develop systems that support communication. In this regard, a possible approach for office environments could be personalized wake-up phrases, which, for example, make it possible to inform a person wearing headphones that anyone wants to talk to them. These must consider the requirements and needs of the addressed person, the addressing person, and the people surrounding them. At the same time, a system must be easily configurable. The resulting discrepancy between variance and accuracy as well as the need to implement the system on small devices is reflected in the general challenges of neural networks. The preliminary evaluation of possible wake-up phrases is a common method to reduce the effort for users and to achieve a satisfactory result as quickly as possible - without knowledge about technical issues or without many field tests until one has achieved a satisfactory result. In our work, we present a first step to close this gap based on AI, which includes a score algorithm for beneficial wake-up phrases. Moreover, we trained different neuronal networks twice based on a self-collected data set by 13 participants for 5 and 4 different Wake-up Phrases for the names Olivia and Charlie and discussed current challenges based on the small amount of data. Therefore, in future research, the data set needs to be expanded concerning the different phrases to achieve more precise results and to improve the corresponding function for determining a score for wake-up phrases. At the same time, the approach must be expanded to include different languages. The current choice of English phrases is primarily based on the availability of audio data sets to support the training of a neural network. The combination of different phrases to activate the system also needs to be further discussed, as different relationships between people in a company may require different forms of greeting. Nevertheless, there were already requirements for wake-up phrases from the current data sets. For example, it was shown that phrases that are too long should be avoided and that this should be considered as part of the ranking. In addition, the current ranking already represents a good basis for further research into personalized wake-up phrases in the context of office work. Finally, the presented approach is a preliminary step toward a possible implementation and requires further research.

ACKNOWLEDGMENTS

The project is funded by the German Federal Ministry of Labor and Social Affairs (BMAS) under the grant number FT 1 – 58330. Thank you to Tobias Nöthlich, Hagen Dähne, Johanna Hoppe, and Jacques-Maurice Walther for their valuable contributions.

REFERENCES

[1] Guoguo Chen, Carolina Parada, and Georg Heigold. 2014. Small-footprint keyword spotting using deep neural networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4087–4091. https://doi.org/10.1109/ICASSP.2014.6854370

[2] I-Fan Chen and Chin-Hui Lee. 2013. A hybrid HMM/DNN approach to keyword spotting of short words. In *Proc. Interspeech 2013*. 1574–1578. https://doi.org/10.21437/Interspeech.2013-397

[3] Jingyong Hou, Yangyang Shi, Mari Ostendorf, Mei-Yuh Hwang, and Lei Xie. 2019. Region Proposal Network Based Small-Footprint Keyword Spotting. *IEEE Signal Processing Letters* 26, 10 (2019), 1471–1475. https://doi.org/10.1109/LSP.2019.2936282

[4] Rajath Kumar, Mike Rodehorst, Joe Wang, Jiacheng Gu, and Brian Kulis. 2020. Building a Robust Word-Level Wakeword Verification Network. In *Proc. Interspeech 2020*. 1972–1976. https://doi.org/10.21437/Interspeech.2020-2018

[5] Hongyi Liu, Apurva Abhyankar, Yuriy Mishchenko, Thibaud Sénéchal, Gengshen Fu, Brian Kulis, Noah D. Stein, Anish Shah, and Shiv Naga Prasad Vitaladevuni. 2020. Metadata-Aware End-to-End Keyword Spotting. In *Proc. Interspeech 2020*. 2282–2286. https://doi.org/10.21437/Interspeech.2020-1262

[6] Jenni Radun, Ville Kontinen, Jukka Keränen, Iida-Kaisa Tervahartiala, and Valtteri Hongisto. 2021. Benefits of Active Noise-Cancelling Headphones in Offices. *The 13th ICBEN Congress on Noise as a Public Health Problem* (2021).

[7] Jenni Radun, Henna Maula, Ville Rajala, Mika Scheinin, and Valtteri Hongisto. 2021. Speech Is Special: The Stress Effects of Speech, Noise, and Silence during Tasks Requiring Concentration. *Indoor Air* 31, 1 (2021), 264–274. https://doi.org/10.1111/ina.12733

[8] J.R. Rohlicek, W. Russell, S. Roukos, and H. Gish. 1989. Continuous hidden Markov modeling for speaker-independent word spotting. In *International Conference on Acoustics, Speech, and Signal Processing,*. 627–630 vol.1. https://doi.org/10.1109/ICASSP.1989.266505

[9] Jan Schmalfuß-Schwarz, David Gollasch, Christin Engel, Meinhardt Branig, and Gerhard Weber. 2024. Open Sesame! Use of Headphones at Work Considering Social Acceptance. In *Computers Helping People with Special Needs*, Klaus Miesenberger, Petr Peňáz, and Makato Kobayashi (Eds.). Springer Nature Switzerland, Cham, 420–429. https://doi.org/10.1007/978-3-031-62849-8_52

[10] Changhao Shan, Junbo Zhang, Yujun Wang, and Lei Xie. 2018. Attention-based End-to-End Models for Small-footprint Keyword Spotting. *CoRR* abs/1803.10916 (2018). http://arxiv.org/abs/1803.10916

[11] Ming Sun, David Snyder, Yixin Gao, Varun Nagaraja, Mike Rodehorst, Sankaran Panchapagesan, Nikko Strom, Spyros Matsoukas, and Shiv Vitaladevuni. 2017. Compressed Time Delay Neural Network for Small-Footprint Keyword Spotting. In *Proc. Interspeech 2017*. 3607–3611. https://doi.org/10.21437/Interspeech.2017-480

[12] A. Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory* 13, 2 (1967), 260–269. https://doi.org/10.1109/TIT.1967.1054010

[13] M. Weintraub. 1993. Keyword-spotting using SRI's DECIPHER large-vocabulary speech-recognition system. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2. 463–466 vol.2. https://doi.org/10.1109/ICASSP.1993.319341

[14] Yundong Zhang, Naveen Suda, Liangzhen Lai, and Vikas Chandra. 2017. Hello Edge: Keyword Spotting on Microcontrollers. *CoRR* abs/1711.07128 (2017). http://arxiv.org/abs/1711.07128