Cross-lingual Voice Activity Detection for Human-Robot Interaction

Nils Höfling [1], Su-Kyoung Kim [2], Elsa Andrea Kirchner [1,2]

1) Medical Technology Systems, Faculty of Engineering, University of Duisburg-Essen, Duisburg, Germany
2) Robotics Innovation Center, German Research Center for Artificial Intelligence, Bremen, Germany

Keywords: voice activity detection, speech recognition, embedded, robot

Abstract

The recognition of language is a two-step process: speech must be recognized as such (1,2) and then the semantics must be understood. For human-robot interaction voice activity detection (VAD) is of great importance (3). Once it is known that a human is talking, speech recognition can be triggered and additional modules in the robot can produce responses to the human, or other robotic behaviors. For online interaction with precise timing especially when using multimodal data (4), it might also be necessary to integrate VAD into a microcontroller or similar embedded system in the robot. Advanced methods exist to enable online and embedded VAD (3). However, some of these methods are trained on biased data, i.e., data in one language, usually English, which can cause problems when used in applications where the interacting human speaks a different language. This is well investigated for speech recognition 5) but poorly for VAD. Language-related issues need to be considered in some applications, such as supporting patients in non-English speaking environments, and may be as important as approaches that handle strong background noise (6, 7).

In this work, we analyze the performance of two different methods to distinguish background noise from spoken words running online on a Raspberry Pi comparing the well-known VAD script of the webRTC standard (8, 9) for real-time communication with a frequency-based approach developed by our group. Both the webRTC and the frequency-based approaches are suitable for online-usage and are independent of an internet connection. We compare latency and accuracy in VAD as a function of language (English and German) and environmental condition for both methods implemented with Python. To import the webRTC VAD-module an open-source python interface (10) was used. The VAD-module is based on a machine-learning model. The basic webRTC VAD often detects short noises as speech. To increase the accuracy of the method a control for the length of the detected signal is added, which greatly decreases the noise identified as speech. As a downside, this may lead to the labeling of some words where only one or two syllables are detected as noise. The frequency-based approach is using signal processing functions of the python library SciPy. Main feature for the feature-based approach is an artificial frequency which is digitally layered on top of the audio signal. The amplitude of the artificial frequency is higher than the amplitude of the audio signal and chosen so that the according peak in the normalized frequency-range only drops if the audio signal has a high enough amplitude but also covers a broad enough part of the spectrum. To increase the accuracy two additional features, which are both based on presence of signals in certain frequency-ranges, are added. Both approaches were implemented on a Raspberry Pi 4B. The onset of voice activity is indicated by a pulse on a GPIO pin of the Raspberry Pi 4B. In future research, this pulse can be sent on a trigger channel, in order to label electroencephalogram data of the interacting human, for example. In this work the pulse is used to evaluate both methods. It is recorded on one audio-

channel while the other channel simultaneously records the speech. The data can then be looked through in an audio program, for example Audacity. For every new spoken statement there should be an according signal on the onset-channel.

The approaches (webRTC VAD, short VAD, and the frequency-based) were tested with four subjects, two females and two males (from 21 to 28 years old). All participants are native German speakers, but also have good English pronunciation. They spoke 24 German and 20 English words each in 3 different environmental conditions: complete silence, background noise, and with echo. For data analysis, we calculated the number of errors (i.e., not recognized, doubly recognized, and recognized as noise) and the number of correct recognitions. We calculated the accuracy of word recognition in percent, because we have a different number of words depending on the language. We considered only two factors for each evaluation, e.g., we compared two methods for each language across three environmental conditions (4 subjects x 3 environmental conditions = 12 samples for each method, see Fig. 1-A). For statistical analysis, we formed Friedman test and Dunn's tests as post-hoc analysis. Bonferroni correction was performed for multiple comparisons.

We found no significant differences in word recognition between the two methods for English words [$p = n.s.$, Fig. 1-A1]. However, the frequency-based method outperformed the VAD method on German words [$p < 0.042$, Fig. 1-A2]. This indicates that the VAD method is less suitable for the recognition of German words. This evidence was supported by further analysis comparing both languages for each method (Fig. 1-B2a, B2b). This further analysis showed that English words were recognized better than German words when the VAD method was used [$p < 0.014$, see Fig. 1-B2a]. However, such language-specific differences were not observed in the proposed method [$p = n.s.$, see Fig. 1-B2b]. These results indicates that the proposed method (Frequency) works very robustly for both languages. Further, we found no significant differences among the three environment conditions when the frequency-based method was used [$p = n.s.$, see Fig. 1-B1b]. This indicates that the proposed method works robustly for all three environmental conditions. However, words in silent environments were much more likely to be detected as noise when the VAD method was used [no echo-VAD (Fig. 1-B1a) in errors (RN) vs. no echo-Frequency (Fig. 1-B1b) in errors (RN): $p < 0.046$].

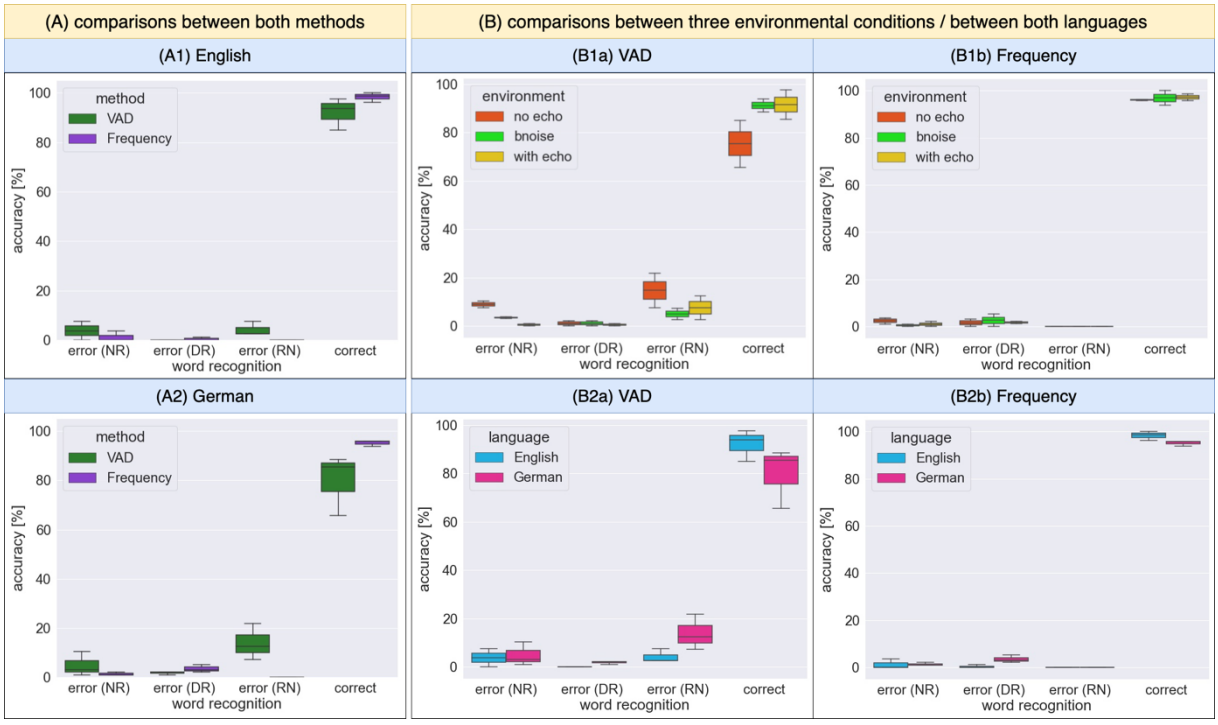

Figure 1 *Accuracy in word recognition for both methods (NR: not recognized, DR: doubly recognized, RN: recognized as noise, no echo: completely silent without echo and background noise, bnoise: with background noise, with echo: with echo, but without background noise)*

Main outcome of this study is that not only speech recognition approaches are language-dependent, but that VAD can also be so. Further, the environment might have a strong influence on VAD which must be considered when transferring approaches from the lab to a real environment. In our current work the developed frequency-based approach is applied to label EEG data for further machine learning based processing in the context of robot-based stroke rehabilitation.

References

(1) Graf, S., Herbig, T., Buck, M. et al. Features for voice activity detection: a comparative analysis. EURASIP J. Adv. Signal Process. 2015, 91 (2015).

(2) Mukherjee, H., Obaidullah, S.M., Santosh, K.C. et al. Line spectral frequency-based features and extreme learning machine for voice activity detection from audio signal. Int J Speech Technol 21, 753–760 (2018). https://doi.org/10.1007/s10772-018-9525-6

(3) Phadke, S., Limaye, R., Verma S. and Subramanian, K., "On design and implementation of an embedded automatic speech recognition system," 17th International Conference on VLSI Design. Proceedings., 2004, pp. 127-132, doi: 10.1109/ICVD.2004.1260914.

(4) Kirchner, E.A., Fairclough, S.H., & Kirchner, F. (2019). Embedded multimodal interfaces in robotics: applications, future trends, and societal implications. The Handbook of Multimodal-Multisensor Interfaces: Language Processing, Software, Commercialization, and Emerging Directions - Volume 3.

(5) Mridha, M.F., Ohi, A.Q., Hamid, M.A. et al. A study on the challenges and opportunities of speech recognition for Bengali language. Artif Intell Rev 55, 3431–3455 (2022). https://doi.org/10.1007/s10462-021-10083-3

(6) Singh, C., Venter, M., Muthu, R.K. et al. DSP-based voice activity detection and background noise reduction. Int J Speech Technol 21, 851–859 (2018). https://doi.org/10.1007/s10772-018-9556-z

(7) Khan, W., Jiang P., and Chan, P., "Word recognition in continuous speech with background noise based on posterior probability measure," 2012 IEEE International Conference on Electro/Information Technology, 2012, pp. 1-7, doi: 10.1109/EIT.2012.6220711.

(8) Available at: https://webrtc.org/

(9) https://www.w3.org/2021/01/pressrelease-webrtc-rec.html.en

(10) https://github.com/wiseman/py-webrtcvad