Command and Conquer: Towards a Robust Inventory of Voice Commands for HMI in Factory Halls

Sara M. Pearsell[1], Oliver Niebuhr[2]
[1] University of Southern Denmark, Sønderborg 6400, E-Mail:pearsell@sdu.dk
[2] University of Southern Denmark, Sønderborg 6400, E-Mail: olni@sdu.dk

Introduction

With the frequency of technology use in daily life rapidly increasing, leveraging and incorporating its various capabilities into an array of areas is crucial. This permits for improvements in simple daily tasks and the advancement of technology used in human-machine-interactions. Voice related technologies, like Automatic Speech Recognition (ASR) systems, are swiftly evolving with these advancements but not uniformly across all fields. In environments with acoustic noise, speech and voice technologies have not made the same impressive technological advancements as other areas.

Industrial work environments, like factory halls, would greatly benefit from the implementation of acoustic controls, including speech command-based inventories. Through utilizing the voice to direct aspects of work such as machinery, workers gain access to physical mobility on the work floor. This could allow for movement between different machines and work areas rather than the stationary demands currently in place where a worker is assigned to a machine in which they must manually interact with a machine for it to operate. For example, pushing a button to start or stop the machine. This would also have a positive effect on physical and cognitive effort from employees while simultaneously increasing efficiency and decreasing costs for employers.

Our research aims to develop and construct an acoustically robust command inventory which is resistant to signal deterioration in loud industrial environments. This research is a part of the Arrowhead project, which aims to create new automated and digitized tools for the European engineering industry [1]. Currently, there is no known usage of, or research on ASR within these loud, industrial environments, in fact there is not even an audio corpus of industry noises for simulation and testing. Other ASR research indicates that many ASR systems struggle with word recognition in environments with general background noise with ASR word accuracy decreasing with noise and/or when the distance between users and input devices increases [2]. Research on speech-in-noise has found speech shaped noise has significant detrimental effects on word recognition scores, and white noise having significant implications but to a lesser degree [3]. This research on other types of noise effects provides a starting point for understanding the potential outcomes of industrial noise.

In addition to understanding and developing effective technology to recognize speech in noisy industrial spaces, sound-based interfaces of this nature also need to function in a strategic way for an array of different users. Specifically, the commands need to be operatable for users from several different language backgrounds. In Germany, for example, Destasis (Germany's Federal Statistical Office) found sectors like metal construction have a workforce consisting of nearly 30% of immigrant workers. Other industrial sectors, like motor vehicle production reveal a similar trend with higher percentages of immigrant workers. This means understanding and integrating common linguistics features from across several languages, such as word and sound structures, and phonemic inventories and phonology, to form more universally approachable commands is necessary [4]. Hence, the combination of linguistic understanding to produce a degree of language universality and effectively utilizing acoustic signal processing to overcome issues of speech recognition in noisy acoustic environments is an important fusion.

The current research is a part of the beginning phases of this exploration. Using linguistic information including observing the overlapping phonology of several languages (Romanian, Hungarian, Bulgarian, German, Danish, Polish, Turkish, Hindi, Urdu, and Farsi) in combination with syllable structure, and the acoustic signal information of sounds including acoustic energy and spectral information, we build an initial word list consisting of 30 different nonwords. We opted for nonwords to ensure there was no semantic interference and a basic CVC (consonant, vowel, consonant) disyllable structure to aid in cross linguistic approachability, ease in learnability, while remaining long enough in duration to ensure recognition to speech recognition systems.

Our first study examines the acoustic impact of various types of noise (i.e., white noise, speech shaped noise, and three different types of industry noise) on word recognition in ASR systems (i.e., Google Translate and Apple Dictation) at increasing dB level of noise. Since our selected ASR systems use real words, we matched 17 real English words which were acoustically similar in both syllable structure and phonemes to a subset of our initial nonword command list. The second study examines the linguistic symbolism of our nonword inventory to 17 most frequently used commands in real industry settings (e.g., *stop*, *start*, *increase speed*, *display*, etc.). This is to observe if there are any intuitive relationships between our nonwords and real word commands aiding in learnability and intuitiveness for ascription of nonword definitions.

To summarize, our two current studies evaluate the accuracy of speech recognition of ASR systems in acoustically noisy industrial environments, as well explore any naturally intuitive relationships between our design nonword command inventory to real word commands in industrial settings. Together, we aim is to develop insights on effective integration between users and speech technologies to increase efficiency, reduce costs, and aid workers overall wellbeing

Methods

For our first experiment evaluating the robustness of our design nonword inventory within noise, we selected 17 real words which matched, as closely as possible (e.g., nonword

would be *sisap* and its real word match, sea salt), to 17 of our created nonword commands in syllable structure and phonemes. We then selected two common and readily accessible ASR systems ASR systems, Google Translate and Apple Dictation and three different types of noise: white noise, speech shaped noise, and three different industry noise. The industry noise was professionally recorded with a high-quality microphone (ZOOM H6 with SSH 6) approximately one meter away from real machinery on industrial floors. Initial visualizations of the acoustic industry noise demonstrated widespread and sporadic acoustic signals across several spectrums in audible signal and with some audio samples containing unique machinery noise and/or interjecting machine reverberations. From these recordings, we selected three machinery noises which were acoustically distinct from one another to ensure a varied sample representative of the different acoustic noise present in real industry settings (see Figure 1).

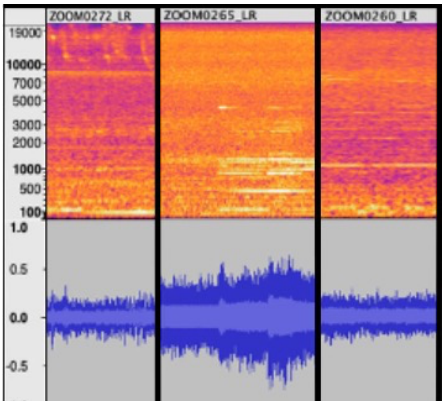

Figure 1: Spectrograms of the three industrial noises uses in the ASR experiment. From left to right, the noises were a grinding machine (Planomat HP Blohm; ZOOM 0272), a cutting and polishing machine (MAM72-70V; ZOOM0265), and a milling and griding machine (Röders TEC RXP500 076DSH; ZOOM0260).

We had two speakers, one male and one female, record each of the 17 real words. By testing the different genders, we aimed to investigate if any effects on word recognition exist due to fundamental frequency (f0) differences or ASR model training [5]. Previous ASR research found the average word error rate for female speakers is much lower than males [6]. It is of interest to find if this result persists in other settings including industrial settings. Since a large percentage of industrial environments employ male workers, understanding this potential implication of decreased work recognition accuracy with the incorporation of ASR systems could create problematic and costly outcomes.

Our second experiment wanted to explore if there were any natural associations with the nonwords to common place commands within the industrial realm. This goes into linguistic symbolism which found nonwords (both acoustic and orthographic) like *kiki* and *bouba*, for example, are consistently and intuitively match to specific descriptions [7, 8, 9, 10]. In the case of *kiki*, the vast major of participants matched kiki with a spikey object whereas *bouba* is matched to a round object. Incorporating the consistent results of the previous research to our nonwords, we aim to understand the intuitive relationship between the nonwords and relationship to real word commands. This could aid in assisting users to learn the nonword commands more easily while being innate in their meaning for a large array of language backgrounds.

Experimental Setup

For our ASR and noise experiment, we set up two sets of loudspeakers. One speaker set (left and right channels) played the different types of noise and were located on either side of one loudspeaker set in the centre. The central loudspeaker played the speech stimuli only (calibrated to 70 dB). Each word from both the male and female speakers were presented at increasing levels of loudness (from 64 dB to 79 dB) as well as with a baseline condition of 0 dB, where no noise was played. A microphone was placed approximately 60 cm in front of the central speaker and was used to capture the audio used for ASR recognition. Apple Dictation (an ASR integrated on Apple devices) and Google Translate (a mobile and web-based ASR) were tested. This was due to each of these systems being widespread in their accessibility. Each word way played three time and give a score of three based on the accuracy of recognition to ensure a clear majority outcome.

For our linguistic-symbolism experiment, ten participants from ten different language backgrounds (Romanian, Hungarian, Bulgarian, German, Danish, Polish, Turkish, Hindi, Urdu, and Farsi) were presented the 30 audio stimuli (recorded by a female speaker) alongside a list of 17 real word commands (e.g., *stop*, *start*, *increase speed*, *decrease speed*, etc.). Participants were instructed to listen to all the audio stimuli and then assign each of the 17 real word, only once, to the nonword they most closely associated with each command. This resulted in 17 commands being assigned to 17 nonwords.

Results

For our ASR experiment, a general mixed-model analysis demonstrated significant effects on word recognition accuracy for noise level, ASR system, and speaker gender. Accuracy decreased in a systematic way when noise increase (from 64 dB to 79 dB, see Figure 2). Apple Dictation consistently outperformed Google Translate across most conditions, except for speech shaped noise (where Google Translate excelled). Contrastively, Apple Dictation performed better in industrial and while noise conditions.

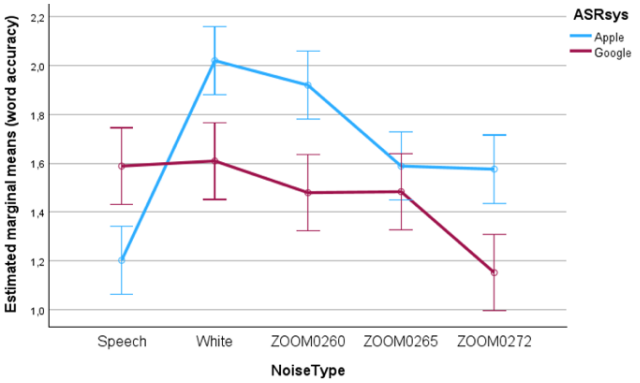

Figure 2: Word accuracy calculated by estimate marginal means for ASR across noise stimuli.

The model also revealed an interaction for gender and ASR, with Apple Dictation producing significantly higher accuracy scores for female voices, while Google had a smaller but

notable advantage for male voices. For noise type, female voices were generally more recognized and more accurately except for the milling and grinding machine (ZOOM0260) where both genders had an equal performance (see Fig. 3).

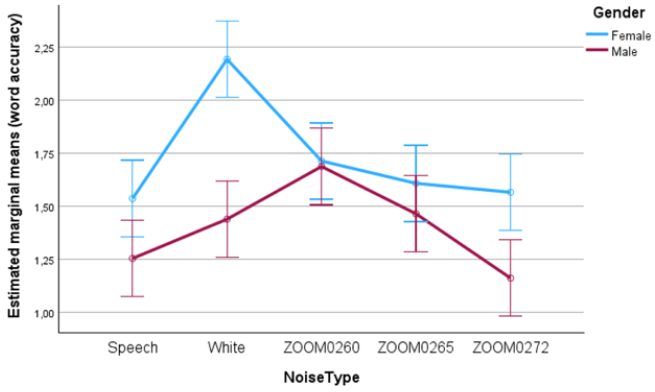

Figure 3: Word accuracy calculated by estimated marginal means for gender across noise stimuli.

We also examined other characteristics of the words themselves and found longer words with higher frequency of use in everyday speech and were also more sonorous in their phonetic structure were more accurately recognized particularly by Google's ASR. Although the intensity of the word had no overall effects, higher intensity provides an advantage at higher noise levels suggesting an interaction of signal strength with noise masking.

For an initial statistical analysis of the linguistic-symbolism experiment, the response data was simplified along three assumptions. First, since our goal is to develop an inventory of machine commands that can be intuitively applied across multiple languages, we initially excluded 'Language' as a statistical factor and examined nonword-to-meaning associations across all listeners, regardless of their linguistic background. Second, as we aim to identify intuitive and systematic associations, we considered only those associations as valid cases in which a nonword was assigned the same meaning at least twice by listeners of a given language. Third, to identify general functional categories of nonwords and to further highlight systematic differences between nonword-to-meaning associations through larger sample sizes, we grouped these meanings into semantic fields.

The five semantic fields applied were: (1) Motion: increase speed, decrease speed, rotate; (2) Activity: start, stop, restart, hold; (3) Adjustment: override, restart, cancel; (4) Direction: left, right, up, down; (5) Interface/Safety: display, mode, emergency.

Initial indications that the 100 participants in the linguistic-symbolism experiment made nonword-meaning associations beyond chance level were provided by a Chi² goodness-of-fit test. This test analyzed how often a nonword was assigned a particular meaning more than twice (across all five semantic fields) and to what extent this frequency distribution deviated significantly from a uniform distribution across all nonwords. The results showed a significant trend indicating a deviation of the empirical frequencies from a theoretical (random) uniform distribution ($\chi^2[29]=39.92$, $p<0.09$). Figure 4 illustrates that 50% of all nonword-to-meaning associations (with ≥2 occurrences) were concentrated in only 10 out of the 30 nonwords. Some nonwords were systematically associated with specific meanings, whereas others were not.

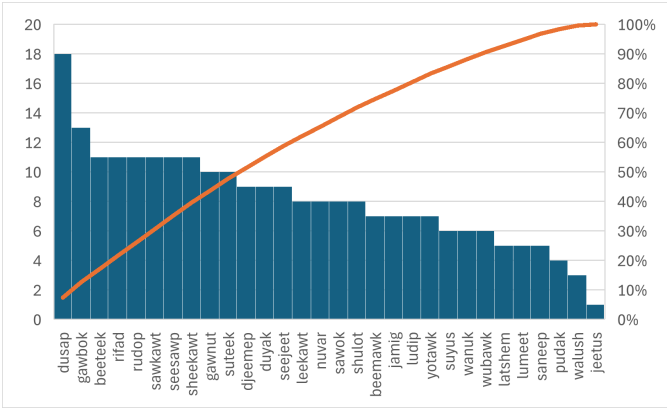

Figure 4: Frequencies (in bars) of nonword-to-meaning associations across the 30 nonwords from most to least frequent and the additive percentage relative to all associations with ≥2 occurrences (N=248).

Further Chi² goodness-of-fit tests examined at the level of the individual nonwords whether their associations with meanings occurred significantly more frequently within specific semantic fields or were uniformly (unspecifically) distributed across all fields. Including significant trends ($p\leq0.1$), 11 nonwords exhibited a systematic and well-defined association with a specific field. Table 1 summarizes these results.

Table 1: Results summary of nonwords that were assigned beyond chance level to a specific semantic field.

NONWORD	$\chi^2[4]$	P	MAIN SEM. FIELD(S)
BEEMAWK	8	0.09	Interface & Safety
DJEEMEP	10.4	0.033	Interface & Safety
DUSAP	14.8	0.005	Activity, Direction
DUYAK	27.1	<0.001	Direction
GAWNUT	11	0.026	Interface & Safety
LUMEET	8	0.09	Motion
LUDIP	8	0.09	Direction
NUVAR	17	0.001	Adjustment
RIFAD	14	0.007	Activity
SEESAWP	16.7	0.002	Activity, Direction
YOTAWK	8	0.09	Direction

Two key findings in Table 1 are particularly noteworthy. First, we can positively conclude that each of the five semantic fields is covered by at least one of the nonwords. This indicates that there is an intuitive, cross-linguistic association for each semantic field with at least one nonword command. Second, we observe that nonwords containing at least one open vowel (e.g., /a/) and obstruent consonants are more frequently associated with meanings related to Adjustment, Interface, and Safety. In contrast, nonwords with sonorant consonants and closed vowels (such as /i,u/) tend to be associated with Motion, Activity, and Direction. This systematic pattern may serve as a valuable foundation for further nonword generation and testing.

Discussion

Our two experiments provide insights into several important aspects regarding both the nonword command list as well as

the effects of noise, specifically in industry, on speech recognition. Our ASR experiment which examines how noise type, noise level, and speaker gender affects word recognition accuracy (in Google Translate and Apple Dictation) reveals certain ASR systems perform better in industrial noise than others. Apple outperforms Google in most conditions, except for speech noise. Apple remains more suited for industrial noise with its ability to adapt in fluctuating and broadband conditions, whereas Google's strength is found in speech noise. The gender matters for ASR systems as well with Apple favoring female voices while Google shows a slight advantage for male speakers. Acoustic and linguistic features of words demonstrate effects with aspects like greater sonority, longer durations improving accuracy.

From the visual interpretation of the linguistic symbolism experiments, there appears to be some nonwords more readily associated with real word commands than others. Some nonwords were not linked to any commands, while some were less consistently linked to real words. Future statistical analysis will provide more insight into these relationships.

Together, these findings highlight the complexities of interplay between ASR type and design, noise type, and speaker attributes as well as the complexities of

Conclusion

These two studies show the importance of noise robust command inventories. Firstly, the need for careful selection of ASR system and for the acoustic and linguistic features of the commands themselves. As our experiments show, certain ASRs, like Apple Dictation, are programmed with great resistance to features within industrial noise while others, like Google Translate, are less optimal and struggle in the same environment. This is also true of the ASRs ability to handle aspects like speaker gender. The findings of the ASR experiments highlight the need for a tailored ASR system, or, more generally, a speech recognition system, to handle the extra external influences of noise in industrial settings. Furthermore, our linguistic symbolism experiment suggests some nonwords align with potential user expectations in aiding in narrowing our larger nonword list down to a more intuitive set of commands.

Future research will incorporate the results of these experiments to further refine the currently proposed list of 30 nonwords. This refined list will incorporate the learnability, and cross-linguistic consistency while simultaneously ensuring the robustness of the nonwords to integrate effectively into industrial HMI interactions. We also aim intend to include expanding ASR evaluations to include additional commercial, and open-source speech recognition systems, investigating real-world implementation scenarios with the industrial noise previously collected from real word industry floors, and refining the phonetic design and linguistic structure of the nonwords to enhance both ASR recognition and user recall. Through the combination of these two experiments, as well as future experiments, we aim to fully develop our command inventory to a level of success which balances machine intelligibility with a user-friendly design, essentially improving and optimizing HMI efficiency for noisy, industrial workplaces.

References

[1] Varga, P., Blomstedt, F., Ferreira, L. L., Eliasson, J., Johansson, M., Delsing, J., & de Soria, I. M. (2017). *Making system of systems interoperable – The core components of the arrowhead framework.* Journal of Network and Computer Applications, 81, 85–95.

[2] Rodrigues, A.; Santos, R.; Abreu, J.; Beça, P.; Almeida, P.; Fernandes, S.: Analyzing the performance of ASR systems: The effects of noise, distance to the device, age and gender. Proc. XX Int. Conf. Human Comput. Interact., 2019, 1–8.

[3] Lee, S. H.; Shim, H. J.; Yoon, S. W.; Lee, K. W.: Effects of various background noises on speech intelligibility of normal hearing subjects. Korean Journal of Otorhinolaryngology-Head and Neck Surgery 52 (2009) 4, 307–311.

[4] ICC Immigration. (2025, March 11). *The growing role of workers with immigration backgrounds in Germany's labor market.* Retrieved April 5, 2025, from https://iccimmigration.ca/the-growing-role-of-workers-with-immigration-backgrounds-in-germanys-labor-market/

[5] Ngueajio, M. K., & Washington, G. (2022, June). *Hey ASR system! Why aren't you more inclusive? Automatic speech recognition systems' bias and proposed bias mitigation techniques. A literature review.* In *International Conference on Human-Computer Interaction* (pp. 421–440). Cham: Springer Nature Switzerland.

[6] Adda-Decker, M.; Lamel, L.: Do speech recognizers prefer female speakers? Proc. Interspeech, 2005, 2205–2208.

[7] Ćwiek, A.; Fuchs, S.; Draxler, C.; Asu, E. L.; Dediu, D.; Hiovain, K.; Kawahara, S.; Koutalidis, S.; Krifka, M.; Lippus, P.; Lupyan, G.: The bouba/kiki effect is robust across cultures and writing systems. Philosophical Transactions of the Royal Society B 377 (2022) 1841, 20200390.

[8] Fort, M.; Schwartz, J.-L.: Resolving the bouba-kiki effect enigma by rooting iconic sound symbolism in physical properties of round and spiky objects. Scientific Reports 12 (2022) 1, 19172. URL: https://www.nature.com/articles/s41598-022-23623-w

[9] Cuskley, C.; Simner, J.; Kirby, S.: Phonological and orthographic influences in the bouba–kiki effect. Psychological Research 81 (2017) 1, 119–130.

[10] De Carolis, L.; Weiss, M.; Fagel, S.; Roussel, N.; Werner, S.; Le Bel, J.: Assessing sound symbolism: Investigating phonetic forms, visual shapes and letter fonts in an implicit bouba-kiki experimental paradigm. PLOS ONE 13 (2018) 12, e0208874. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0208874