The Benefits of Contextual Information for Speech Recognition Systems

1st Martin W. Kinch
Department of Engineering Science
University of Greenwich
Chatham Maritime, United Kingdom
M.W.Kinch@greenwich.ac.uk

2nd Wim J.C. Melis
Department of Engineering Science
University of Greenwich
Chatham Maritime, United Kingdom
Wim.J.C.Melis@greenwich.ac.uk

3rd Simeon Keates
Department of Engineering Science
University of Greenwich
Chatham Maritime, United Kingdom
S.Keates@greenwich.ac.uk

***Abstract*—This paper demonstrates the significance of using contextual information in machine learning and speech recognition. While the benefits of contextual information in human communication are widely known, their significance is rarely explored or discussed with a view to their potential for improving speech recognition accuracy. The presented research primarily focuses on an undertaken empirical study that looks at how context affects human communication and understanding. During the study, comparisons between human communication with and without context, have shown overall recognition improvements of over 30% when contextual information is provided. The study has also investigated the importance of the former/middle/latter part of a word towards recognition. These results show that the first two-thirds of a spoken word are key for humans to correctly infer a word. The conclusions from the performed study are then drawn upon to identify useful types of context that can help a machine's understanding, and how such contextual information can be gathered in speech recognition and machine learning systems. This paper shows that context is not only highly important for human communication, but can easily provide a wealth of information to enhance computational systems.**

***Index Terms*—Machine Learning, Contextual Information, Speech Recognition, Natural Language Processing, Context-Aware Computing, Artificial Intelligence**

I. Introduction

Human communication is normally supported by additional contextual information that aids understanding. Context itself, as described by the Oxford dictionary, is: ”The circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood.” [1].

A lack of, or misinterpretation of context often results in misunderstanding, which then tends to negatively impact human communication. Despite this and in contrast to human communication, which relies quite extensively on context, Machine Learning (ML) algorithms only support contextual information processing to a limited extent within their design.

This research aims to explore how context affects human communication, with a view to how ML can learn to process context in a similar manner to humans. Tools that use ML, such as Natural Language Processing (NLP) and Speech Recognition (SR), are expected to benefit from processing context, specifically by improve recognition rates. Furthermore, an investigation into the background of context, in both computational and human research has also been undertaken, which helps to support the design and rationale of the contextual approach.

II. Background

While context has previously been used in computing fields, the types and variety of context are limited, often relying purely on location data. However, humans often use a wide range of contextual information, from location all the way to a persons emotional state. Clearly, the full range of context is not always used in human communication, but having a wider range than what is often used in computing, is beneficial for communication. Support for this can be seen in fields such as psychology, which is where the background review starts before moving into how context is currently used in computing.

A. Context from a Psychology perspective

An area where humans benefit strongly from context is memory recall, specifically in a topic known as context dependent memory. Context dependent memory is the process by which human memories become linked with external stimuli, for instance location, environment or a number of other factors. Research has shown that humans remember details significantly better when the related context is similar to when a task is first performed [2], [13].

Furthermore, research has also shown that exposing someone to background noise, which has a connection to a performed task, can improve a persons proficiency in that task and aid memory recall [12]. As such, it is not unreasonable to suggest that contextual information aids memory recall and that it shows that humans not only derive/use context in communication, but also in how they store information within the brain, which indicates its overall significance.

Moreover, further research indicates that adults with a limited reading ability often use contextual information present in surrounding text, to make inferences as to the meaning of sections they cannot understand [5]. An ability that has implications for spoken languages as well, for when a person is not fully aware of words being spoken (such as not hearing something correctly) they can often still understand the intention, if not always the full content, of the communication.

Clearly, contextual information, such as sounds, situations and communication style, allow humans to infer certain details that do not need to be explicitly stated, such as the mood of a person. Consequently, context must also improves a person's ability to make judgments by inferring important information, thus ”filling in the blanks” that can occur during communication. Utilisation of this form of contextual processing could have a plethora of benefits of NLP/SR and ML implementations.

B. Context within Computing

While contextual information has already been used in the field of context aware computing, the types and variety of contexts used is limited. For example, some tools keep track of an individual's location to provide them with personal information, such as phone calls, by routing this information/call to the nearest available work-station [8], [14].

Other means of obtaining contextual information can be derived from accelerometer or light sensor output, as currently found in mobile devices [9]. In this case, the obtained information is used to determine the correct orientation of images on a screen or to automatically adjust brightness of the screen, and so the contextual information is used to adapt processes that relate to the context/mode of operation. Alternatively, accelerometers have also been used as an authentication method for mobile devices [7].

Considering that context is all about the situation someone/something is presented with, it is this type of input that would be beneficial to ML. This could be obtained through images and/or sounds and has been demonstrated through, for example; guidebooks that recognise artwork, to help tailor a museum visitor's experience [11], but also in using video to improve speech recognition rates [3].

Events are also a useful form of providing contextual input, for example, patents filed by [6] demonstrate how a car crash can be used to trigger responses from a cars' control systems. While context in sentences has also been used in Hidden Markov Models and Deep Neural Networks to aid speech recognition by determining individual utterances [4], [10].

As seen in the previous examples, very few systems utilise multiple types of contextual information, such as time, location, mood, calendar information or the wide range of contextual information stored on the average smart phone, to identify or benefit from a context. Mostly, they use one or two contexts for a specific purpose, but not as a factor in system learning or training, where the connections between communication and context could be beneficial, such as in SR or NLP. Yet, there is the potential that by having these contextual cues that both SR and NLP recognition rates could be improved, due to an improved understanding of the related contexts.

III. Methodology

To evaluate the effects of context within human communication, a survey was created and made available to participants through an online survey tool. The survey was conducted online so that it could attract a diverse audience worldwide, ensuring results are as representative of the population as possible. Having results that are representative in this fashion, helps to ensure that any potential demographic bias would be identified, should it exist.

The survey consists of several questions covering: context within text, context within audio and how incomplete audio is best understood. The design of this survey has taken inspiration, in part from [15].

A. Text with Context

The first set of tests aims to explore the difference in recognition accuracy for sentences with one or multiple types of context. In these tests, the context consisting of multiple contextual types was text (due to the large amount of context present in even simple sentences), and the singular type of context was audio. The text included statements relating to a situation, person(s) or actions, while the audio only related to a locational/situational clue e.g. background noise.

The text based test is comprised of three questions, which are sentences with two omissions. These questions each consist of four multiple-choice answers. For example, one of the sentences, is: ”Mark walked down to his local [omitted word]. He was hoping to pick up some food for his [omitted word] tonight.”. Options to fill the space are: ”Bank-Dog”, ”Council-Meeting”, ”Shop-Dinner” or ”Beach-Party”. Questions are designed so that only one of the available choices could be considered correct. The use of two-part answers, provided extra context, so that the participants could further match contexts (in addition to those found in the question) thus aiding the selection of the correct answer. The audio part of this test is explained further in the next subsection.

B. Context for Auditory Information

The tests in this section, considered the importance of context within audio, as well as comparing the difference between audio with and without context. Each test had three questions, where one word was obscured by noise. In the cases with context, background sounds were added to provide a contextual clue, while the cases without context had no such sounds. In practice, each audio sample contained a single sentence: ”The next departure will be at [omitted word] 4”, where the omitted word is either ”platform” or ”gate” relating directly to a train station or airport setting.

Participants were asked at the start to indicate how familiar they were with the various modes of transport, thus reducing potential bias based on previous knowledge of the presented context. During the test, the participants had the freedom to play the sounds at their leisure, while they were asked to choose from the following options: ”Gate”, ”Platform” and ”Not Sure” always appearing in that particular order.

C. Word Part Significance

Further audio related tests aimed to explore the importance of each part of a single word. For which, several words were split into three parts, and these were respectively obscured at either the front, middle or back. Overall, six words were used in this test, of which there was one ”pair” per word-part. Each word was only used once to ensure that memory or prediction would not influence the results.

Each pair consisted of one word with two syllables e.g. ”Mother”, and one with three syllables e.g. ”Daffodil”. Words are provided as audio samples and there was no order in the way they were presented to the participant with regards to number of syllables and/or part of word that had been obscured. Participants were asked to provide answers in a free text box. Which, were graded as 1 for a correct answer, 0.5 for a partially correct answer (e.g. sounding similar or being close enough to the expected answer) and 0 for an incorrect answer.

IV. Results

Results from the study are presented below, and where appropriate subjected to a Student t-test. The demographic section details the demographic spread of participants in the survey.

A. Demographics

The study comprised of 80 participants, with a gender split of: 38 females, 28 males, 13 unidentified and 1 non-binary. The average age was 44, and ranged from 22 to 79. At the start of the tests, each participant was asked for details pertaining their hearing ability through an online hearing test. Further introductory questions asked for their familiarity using various modes of transport (Train, Plane, Boat, Bike, Bus, Taxi), their first language, nationality, profession and ethnicity. These question were asked so that comparisons could be run with specific demographics focuses, to ensure no particular bias affects the results. The related comparisons found no particular bias in the results relating to demographic information.

B. Text and Audio with Context

Figure 1 details a comparison between text and audio based context. In this example the text based context has multiple types of context, while the audio based context only has one type of context.

As seen in Figure 1 the text with context has an accuracy of nearly 100%, illustrating that the text based context has sufficient contextual information to decipher the correct answer.

On the contrary, audio with context averages to only about 68%, which shows that limiting the amount of context available, significantly lowers the response accuracy. Clearly, having a wider variety and amount of contextual information improves the overall response accuracy, in this case by over 30%. Additionally, when subjected to a t-test the results show a probability distribution of 5e-10, which indicates that both data sets are statistically different, or that more context results in better recognition.

A noteworthy point, is that although text based context had a higher accuracy than audio based context, the main difference (other than information type) was the additional context in the text compared to audio. Thus audio with the same level of context, as the text based context, is likely to improve by similar amounts when provided alongside more types of context.

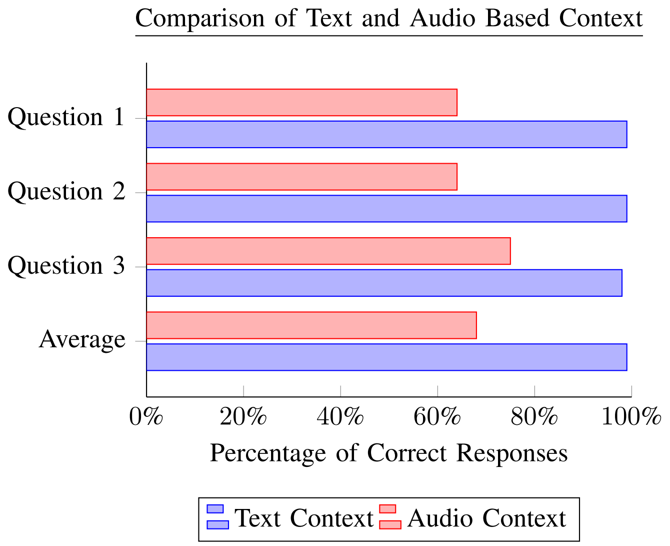

Fig. 1. Comparison between audio with context (one type of context) and text with context (multiple types of context)

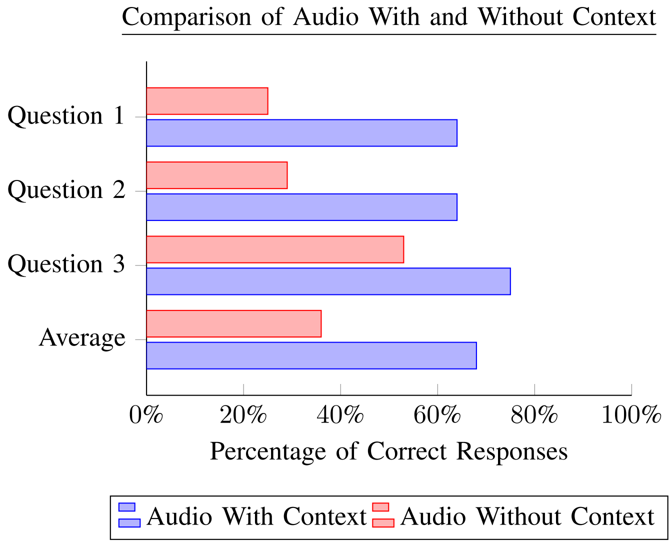

Fig. 2. Comparison between questions with/without context

The outcome of this comparison is as one would expect. The contextual information that is present in the text based test is significantly more dense (more of it) than what was present in the audio based test. This shows that having more context, as long as its accurate, is going to improve recognition accuracy, this is especially true when there is missing or unclear sections of audio. This feature is specifically useful in fields such as speech recognition, where words or phrases can be misspoken or damaged by noise.

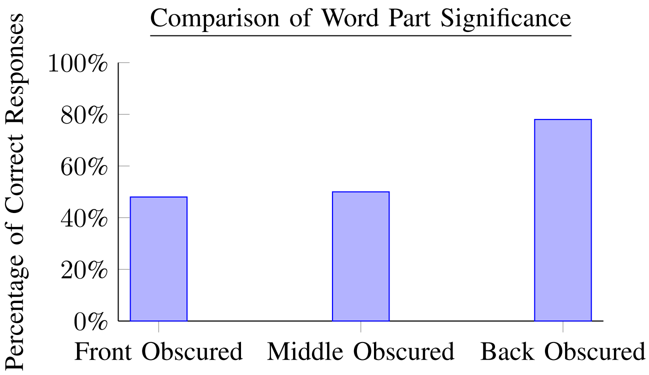

Fig. 3. Comparison of which part of a word is most significant

C. Audio with and without Context

The comparison between audio with context and audio without context, is shown in Figure 2 and details the accuracy of participant responses. The results in Figure 2 illustrate that accuracy is considerably higher for the questions with context, than those without. In this case the audio with context shows improvements of over 30%. Clearly the lack of context severally weakens the accuracy of a participant's response, while adding just one type of context can cause considerable improvements.

Interestingly, there is a slight peak at question three for both audio with and without context. The reason for this seems to be due to popular choice (due to perhaps familiarity) for an answer, which was ”Platform”. ”Platform” was the correct answer for question three and incorrect for one and two, thus given the peak it would seem that those who selected the most familiar option went with ”Platform” including on question three. Thus explaining this peak.

A combined analysis of audio with/without context, subjected to a t-test shows a probability score of 3e-10; once again indicating that both data sets are statistically different, and consequently supporting the hypothesis that context improves human recognition rates. The results are as one would expect, based on the results from the previous comparison in 1. Interestingly the accuracy increase is the same for one form of context and multiple forms of context. Thus it can be considered that using context in multiple forms could constitute an increase to over 60% accuracy against no context at all. However, given that one form of context and multiple contexts are both 30% accuracy increases, this could suggest that accuracy growth becomes slower as more contexts are added. Hinting at a potential peak number of contexts for effective contextual processing.

D. Word Part Significance

Figure 3 shows the results of the word part significance test. This test aimed to identify the parts of a word which are most important for understanding/recognising a whole word. In the test, words were either obscured at the front, middle or back. From the results, it is clear that the highest recognition accuracy of the words, is achieved when the back/latter part of the word is obscured. Whereas, on the contrary, recognition accuracy is poorer when the front or middle part are obscured. This indicates that the most significant part of a word lies in the first parts, and that if a person misses the last part of a word, inferring the correct word is more viable. This can also be noticed in practice where in various languages and/or dialects the endings of words may be ”swallowed”/not clearly pronounced, which often does not significantly affect the understanding.

Although, useful to know that the first parts of a word are most important, this does not necessarily help with reconstruction, as audio can be provided to a system in any condition. However, it does show that predictions made on words that are missing the front parts may benefit more from contextual processing (as it “adds” more data) compared to words which are only missing later sections (as these should be easier, thus safer to infer).

V. The Benefits of Context for Machine Learning

The results presented in this paper show that context is an incredibly important feature of human communication and understanding. However, few ML algorithms process context in the same manner, often only being context-aware in a limited sense (Perhaps only knowing who is talking, rather than a detailed understanding of the current situation). The expectation from research into context is that ML algorithms will have greater detail with which to make inferences. For instance, in speech recognition, rather than dealing with only the recorded words, other details such as: the location, other people present or the persons mood, could all provide indications to the situation being experienced by the speaker. This experienced situation will likely change the likelihood of certain phrases or words being spoken, making the situation and the context that forms the setting for that situation, important factors. By processing these factors alongside the spoken words, not only will predictions about these words be more reliable, but situations where words are unclear or misinterpreted should be somewhat less impactful, as the ML algorithm can make a more informed guess about the “missing” word, than a traditional SR tool.

Despite the benefits that context offers ML, there is a drawback. Since using context will require extra data such as: names, places or any similar detail, this means that the data provided to a ML will not only be considerably larger but also that training could be significantly longer depending on the amount of context used. The expectation is that the improved accuracy will offset these initial training and data generation costs, however management of necessary amount of context should keep the increase in training and data needs to a minimum. For example, is it beneficial to know that someone is in a restaurant? Potentially yes, but does it matter that the restaurant is Italian or Chinese cuisine, probably less important, on average.

A further advantage to context is how easily it can be gather in a real world scenario. Often, people carry smart phones on their person, something becoming more prevalent, as smart phones become a part of daily life, being used to keep in touch, make payments and access the internet. These devices often gather a wide range of data on the user and user habits, but most importantly the data needed for contextual processing is quite specific and can be gathered easily e.g. GPS (Global Positioning Satellite), Message data, Calendar or Meeting information to name a few. These forms of data can be processed to form an overall understanding of the context without requiring complex analysis of user behaviour.

Importantly, using context needs to be as close as possible to human ability when used in ML. The reason for this stems from the speed at which humans can process context, obviously when people go about their daily lives the context is almost omnipresent. It is easy for someone to know what they are doing and why, without much need for previous thought. Thus when using context with ML it is important to keep the processing speed in line with what has come to be expected with ML algorithms in the last few years. The way to achieve this comparability lies in keeping the context linked intrinsically with the task specific data (speech, images or similar), so that the process is not two part; processing task specific date then context. But instead, processing task specific data alongside its contextual equivalent.

VI. Applicability to Speech Recognition

A natural progression for context, especially given its effect on human communication, is using context alongside SR. Moreover, since most SR systems are now based on ML methods and algorithms, the same advantages and disadvantages from the previous section are still valid. However, SR does have some unique considerations, namely the linkage between phrases/words and a context, as well as the potential to identify a context from an audio stream.

The linkage between a context and phrases/words will, as mentioned previously, still need a mixture of task specific and contextual data for training. It is expected this form of training would result in a “contextual model” that could relate, perhaps to start, a few contexts with certain phrases and words which would not only increase accuracy, but could adapt as it is trained further in real life.

While not guaranteed to provide clear data, background noise such as cars, trains or even classrooms also have the potential to inform the SR of the current context. As shown by the research in this paper, background noises can dramatically change a persons perception of an event. If the background noise can be utilised this offers another form of context that SR can exploit.

The overall expectation is that when words/phrases that are linked to a context, the SR system will have an extra factor (context) to strengthen its certainty of a certain word or phrase being correct. It will also allow ”odd” words that do not fit the context to be reconsidered, thus limiting when an SR system offers an unusual word, rather than a more reasonable choice.

VII. Discussion

This paper has shown that context has real potential to improve the accuracy rates of ML and SR systems. Not only does context offer ML and SR with a wider range of relevant input data, but it also brings the systems inference in line with how humans currently deal with and process communication. However, to offer a fair appraisal it is important to note there are limitations to how context functions and there is potential for context to be expensive to process. It is also important to ensure that context is accurate and that it can be correctly recognised.

The main issue with context is the data burden it can generate, ML system are normally designed to reduce dimensionality in favour of principal variables to improve processing speed. Adding context will mean adding more principal variables (or one if a separate recognition stage is used) which will increase processing speeds. This could cause problems in performance critical situations. The question then becomes does the system need to be fast, or accurate, which will heavily influence whether context should or should not be used. However, it is expected that context would only increase processing costs by a small to moderate amount depending on the overall process. Thus if an application is not time critical - such as search - context may well be very appropriate.

Context also needs to be correctly identified. While this is easy enough to ensure when collecting data from a smart device, processing background noise may prove difficult to fit into a context. The important steps here are to ensure the weighting of the context is appropriate to the application and does not provide false negatives. It may also be useful to weight each source of context differently depending on its reliability.

Overall, using context does entail some risks and could take somewhat longer to train and process than traditional SR/ML systems. However, accuracy is likely to improve especially in situations where there is either high noise or lost/damaged data. Which, on balance is beneficial in applications that require accuracy over processing speed or training speed.

VIII. Conceptual Model

To further illustrate how context could be used in SR a block diagram can be seen in figure 4. This diagram shows how a conceptual context processing SR could function. In this example there are three sources of contextual information: location data, calendar data and message data. These data streams will be used to determine the overall context e.g. in a meeting or waiting for a bus. Once the overall context has been established it will be combined with the input speech data and passed to the SR. The SR, being influenced by the context, will process the speech data and provide output; as the text version of the input speech data. Of course, there is more than one potential method for processing context. However, this simple example helps to show how a contextual processing algorithm would flow and what the expected output would be.

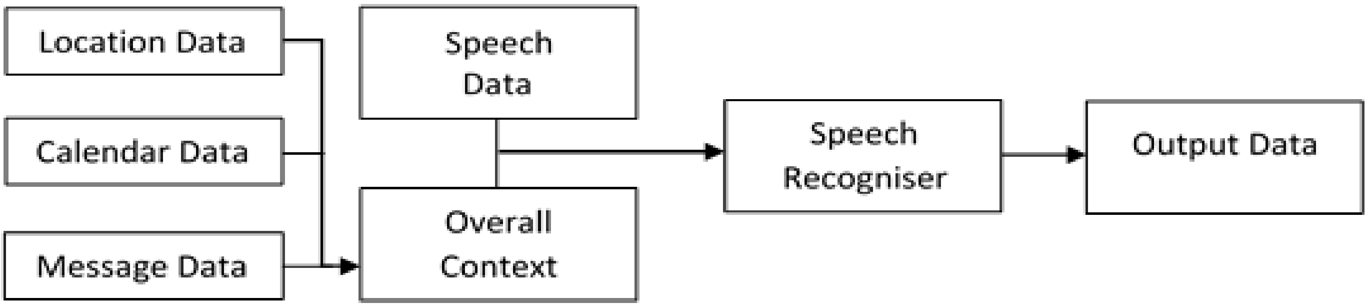

Fig. 4. A block diagram showing how context could be utilised

IX. Conclusion

This paper has evaluated the benefits and drawbacks of using context in ML/SR. Overall, context has the potential to improve ML/SR accuracy rates, but may also have further benefits identified as research continues. Background research has shown that humans use context considerably, especially during communication and memory recall. While context-aware computing has processed context for decades, however it has rarely done so in the manner described in this paper.

A study has also been undertaken that looked at how context impacts human communication. The study has shown that not only does having context, compared to no context, improve recognition accuracy by over 30%. But also that have more than one type of context can further improve accuracy again by 30%. Showing over 60% improvement between having no context and having multiple types of context. This study helps to support context as an important factor for human communication and potentially for ML/SR applications.

The study also highlight the fact that the first two-thirds of a word are the most important when trying to recognise a partially obscured word. This shows, at least in the English language, that ”reconstructing” a word is much easier when the former parts of the word are heard, and significantly harder when they are obscured.

Despite the fact that using context in ML and SR could cause longer learning or processing times, there is a strong argument for its use in accuracy dependent applications, or in those afflicted by high noise. It is also expected that these issues will reduce with future research, when more efficient methods for processing context can be devised. Overall, this paper makes the case that context holds considerable depth and potential for further research, and that its use could help improve ML and SR by creating a more human-centric approach to their learning and processing.

X. Further Work

Further work will look to explore the types of context that are beneficial to SR systems and how a context can be identified and used to improve its respective recognition rates. To test this theory a simple SR system will be developed to recognise/process both speech and context, with a view to training a relationship between the speech and context using ML. It is expected that by training the SR system in this fashion, future SR accuracy will be improved upon the identification and use of contextual information.

XI. Acknowledgments

The authors wish to thank Josh Davis for his help with distributing the context survey.

References

[1] Oxford University Press, 2017, Oxford Living Dictionary.

[2] D. R. Godden and A. D. Baddeley, “Context-dependent memory in two natural environments: On land and underwater,” *British Journal of psychology*, pp. 325–331, 1975.

[3] A. Gupta, Y. Miao, L. Neves, and F. Metze, “Visual features for context-aware speech recognition,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5020–5024.

[4] S. Kanthak and H. Ney, “Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition,” in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2002, pp. I–845.

[5] S. Ng, B. R. Payne, E. A. Stine-Morrow, and K. D. Federmeier, “How struggling adult readers use contextual information when comprehending speech: Evidence from event-related potentials,” *International Journal of Psychophysiology*, pp. 1 – 9, 2018.

[6] J. Palmer, M. Freitas, D. A. Deninger, D. Forney, S. Sljivar, A. Vaidya, and J. Griswold, “Systems and method to trigger vehicle events based on contextual information,” Feb. 1 2018, uS Patent App. 15/727,227.

[7] A. Primo, V. V. Phoha, R. Kumar, and A. Serwadda, “Context-aware active authentication using smartphone accelerometer measurements,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2014, pp. 98–105.

[8] B. N. Schilit, D. M. Hilbert, and J. Trevor, “Context-aware communication,” *IEEE Wireless Communications*, pp. 46–54, 2002.

[9] A. Schmidt, M. Beigl, and H.-W. Gellersen, “There is more to context than location,” *Computers & Graphics*, pp. 893–901, 1999.

[10] F. Seide, G. Li, X. Chen, and D. Yu, “Feature engineering in context-dependent deep neural networks for conversational speech transcription,” in *2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2011, pp. 24–29.

[11] L. Seidenari, C. Baecchi, T. Uricchio, A. Ferracani, M. Bertini, and A. D. Bimbo, “Deep artwork detection and retrieval for automatic context-aware audio guides,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, p. 35, 2017.

[12] S. M. Smith, “Background music and context-dependent memory,” *The American Journal of Psychology*, pp. 591–603, 1985.

[13] S. M. Smith and E. Vela, “Environmental context-dependent memory: A review and meta-analysis,” *Psychonomic bulletin & review*, pp. 203–220, 2001.

[14] R. Want, A. Hopper, V. Falcao, and J. Gibbons, “The active badge location system,” *ACM Transactions on Information Systems (TOIS)*, pp. 91–102, 1992.

[15] R. M. Warren and R. P. Warren, *Auditory illusions and confusions*. WH Freeman, 1970.