The Benefits of Contextual Information for Speech Recognition Systems

Martin W. Kinch, Wim J.C. Melis and Simeon Keates
Department of Engineering Science
University of Greenwich
Chatham Maritime, United Kingdom
M.W.Kinch@gre.ac.uk, Wim.J.C.Melis@gre.ac.uk, S.Keates@gre.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 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, 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 learn how ML can process context in a similar manner to humans. It is expected that processing contextual information will improve recognition rates of Natural Language Processing (NLP) and Speech Recognition (SR). 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, it is often limited to location data or a singular contextual type. However, humans often use a wide range of contextual information including location, all the way to a person's emotional state. The full range of context is not always used in human communication, but having a wider range than what is currently used in computing, could be beneficial towards human-computer communication. Support for this can be seen in 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 was first performed [2], [3].

Furthermore, research has also shown that exposing someone to background noise, which has a connection to a performed task, can improve a person's proficiency in that task and aid memory recall [4]. As such, it is not unreasonable to suggest that contextual information aids memory recall. It also 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 from the surrounding text, to make inferences as to the meaning of sections they struggle to understand [5]; something that applies even to those with good reading skills. This ability also has implications for spoken languages as well, for example when a person is not fully aware of words being spoken (such as not hearing something correctly) they can often still understand the intention of the communication.

B. Context within Computing

The types and variety of context used in context-aware computing has so far been 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 call to the nearest available work-station [6], [7].

Other means of obtaining contextual information can be derived from an accelerometer or light sensor output, as currently found in mobile devices [8]. 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. Thus, the sensory information is used to adapt the mode of operation. Alternatively, accelerometers have also been used as an authentication method for mobile devices [9].

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. Context 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 [10], but also in using video to improve SR rates [11].

A Patent filed by [12] also demonstrate how vehicular system events can be triggered by contextual information, which is gathered via sensors external to the vehicle. While context in sentences has also been used in Hidden Markov Models and Deep Neural Networks to aid SR by determining individual utterances [13], [14].

As described 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 context. Current systems often 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.

III. Methodology

To evaluate the effects of context within human communication, a survey was created and made available through an online survey tool. Performing the survey online allowed for a diverse audience worldwide to take part, ensuring the results are as representative of the population as possible. Moreover, having a diverse survey helps to ensure that any demographic bias can be identified, if it exists.

The survey consists of several questions covering: singular and multiple types of contextual information, context for auditory information and word part significance. The design of this survey has taken inspiration, in part, from [15].

A. Singular and Multiple Types of Contextual Information

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, a text based test provided the multiple contextual types, due to the large amount of context present in even simple sentences, and the singular type of context was provided via 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, thus aiding the selection of the correct answer.

B. Context for Auditory Information

The tests in this section, considered the importance of context within audio, as well as 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 the same order to avoid participants inferring a pattern.

C. Word Part Significance

Further audio related tests aimed to explore the importance of each part of a single word. For which, various 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 with two words per the front, middle and back of a word. Further more, each "group" of words had one word with two syllables and the other with three syllables. Each word was used once to ensure that memory or prediction did 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 in random order, with regards to the 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 to the expected answer) and 0 for an incorrect answer.

IV. Results

The results from the study are detailed below, with each set of results being subjected to a Student t-test, where necessary. Demographic information has also been gathered and analysed to limit the effect of any potential biases in the data.

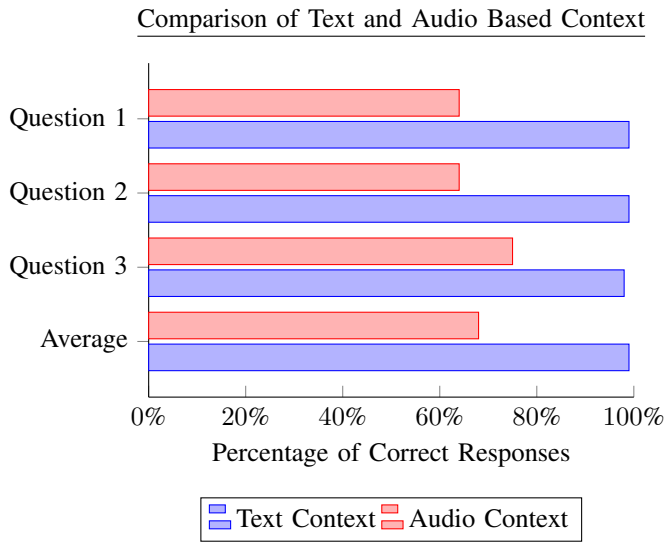

Fig. 1. Comparison between audio with context (one type of context) and text with context (multiple types of context)

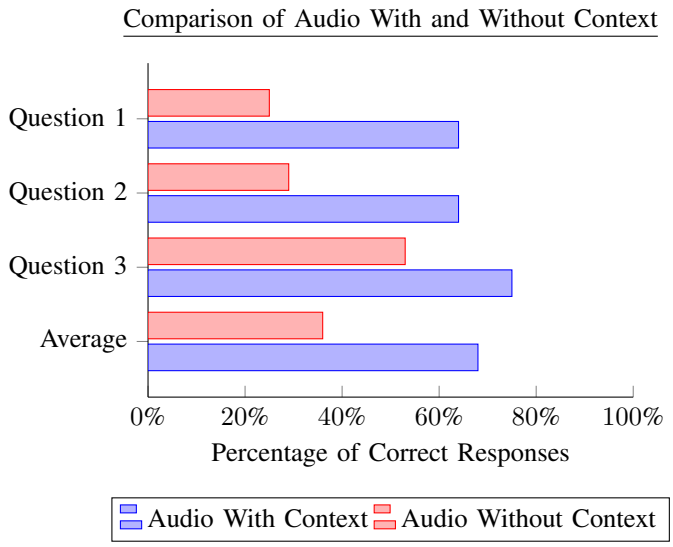

Fig. 2. Comparison between questions with/without context

A. Demographics

The study comprised 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 about their first language, nationality, profession, ethnicity and their familiarity using various modes of transport (Train, Plane, Boat, Bike, Bus, Taxi). These questions were asked so that comparisons could be run with specific demographic 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, with the text based context possessing multiple types of context and the audio only a singular type of context. The results show that text with context has an accuracy of nearly 100%, indicating that the text based context has sufficient contextual information to decipher the correct answer.

Conversely, audio with context averages to 68%, which shows that limiting the amount of context, significantly lowers the response accuracy. Clearly, having a wider variety and amount of contextual information improves the overall response accuracy 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 contexts has a higher accuracy than audio based context, the main difference lies in the additional available context in the text in comparison to the audio. Thus, having audio with the same level of context as the text based context, is likely to improve accuracy equally.

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

C. Audio with and without Context

The results of audio with and without context, is shown in Figure 2. The results in Figure 2 illustrate that accuracy is considerably higher for the questions with context, than those without. Thus having context shows improvements of over 30%. Clearly the lack of context severely 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 for the answer: "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", which appeared in the middle of the list of options.

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. 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 of over 60%

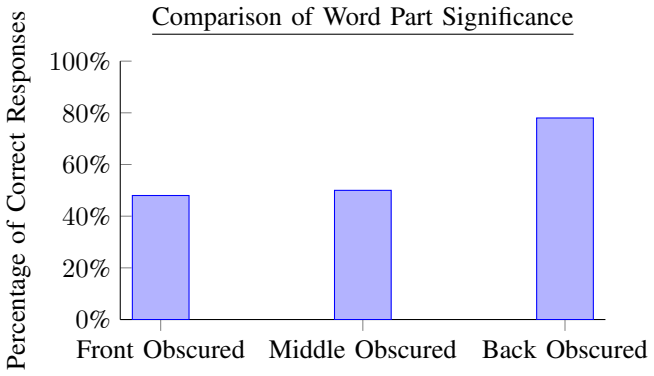

Fig. 3. Comparison of which part of a word is most significant

against no context. However, given that one form of context and multiple contexts both increase accuracy by 30%, this could suggest that accuracy growth becomes slower as more context is added. Indicating that there may be a threshold to 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 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 is achieved when the back/latter part of the word is obscured. Whereas, 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 not be clearly pronounced, which rarely affects the understanding.

While it is 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 shows that predictions made for 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 latter sections.

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 very limited sense. Thus, using context alongside ML will make situations easier to infer and improve solution accuracy accordingly.

For instance, in SR, rather than dealing with just the recorded words, other details such as: location, other people's presence or the persons mood, could all provide indications to the situation being experienced by the speaker. Each situation will have different likelihoods for words and phrases appearing, which can then be used to improve estimates about what is being said. This makes the situation and related data very 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 context-aware ML algorithm can make a more informed guess about the "missing" word, than a traditional SR tool would.

Despite the benefits that context offers to ML, there is a drawback, which lies in the extra data required to function, such as: names, places or any similar detail. This means that the data provided to the ML will become considerably larger but also that training could take significantly longer. The expectation is that the improved accuracy will offset these initial training, data generation and storage costs. However, proper management of the required contexts will limit the increase in training and data costs. For example, is it beneficial to know that someone is currently writing? Potentially yes, but does it matter if they are using a pen or a pencil, probably not.

An advantage to context is how easily it can be gathered in a real world scenario. Often, people carry smart phones on their person that can gather a wide range of data on its use and habits. However, more 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 just a few.

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, context is omnipresent. It is easy enough for someone to know what they are doing and why, without much need for previous thought but the same is not currently true for a machine. To achieve this comparability the context will need to be linked with the input data, forming a relationship between the two. Thus, when new data enters the system with an associated context. The context and the data are evaluated at the same time as part of the same process, which will save computation time.

VI. Applicability to Speech Recognition

A natural progression for context, especially given its effect on human communication, is to use 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 relationship between a context and phrases/words will still need a mixture of task specific and contextual data for training. It is expected that this form of training would result in a "contextual model" that could relate to a few contexts with certain phrases and words, which would not only increase accuracy, but could adapt to new situations as it gets trained in real life scenarios.

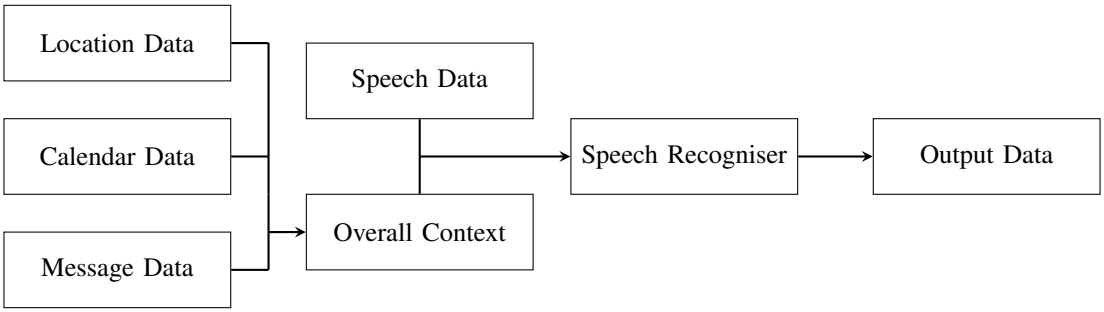

Fig. 4. A block diagram showing how context could be utilised

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 in this paper, background noises can dramatically change a person's 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 are linked to a context, the SR system will have an extra factor to strengthen its certainty of the recognised word or phrase being correct. Unusual words that do not fit the context are also much less likely to occur due to different weightings in the SR system after training.

VII. Discussion

This paper has shown that context has real potential to improve the accuracy rates of 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 the overhead related to processing context, to be considerable. It is also important to ensure that context is accurately identified.

Due to the overhead for processing context there may be limitations to its use 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 if dimensionality of the data is reduced. Thus if an application is not time critical and relies on accuracy, 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 is more challenging. The important steps here are to ensure that context recognition is weighted to ensure that a context is highly likely before suggesting it to the system. Furthermore, less reliable sources of context can be indicated differently so that SR weighting can take this into account when making predictions.

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 levels of noise or lost/damaged data.

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 text 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.

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. Background research has shown that humans use context considerably, especially during communication and memory recall, but while context-aware computing has processed context for decades, the types and variety of contextual information used has been limited.

A study has been undertaken that looked at how context impacts human communication. The study has shown that not only does having context, compared to not having context, improves recognition accuracy by over 30%. But also that having more than one type of context can improve accuracy by an additional 30%. Showing over 60% improvement between having no context and having multiple types of context. The study also highlights the fact that the first two-thirds of a word are the most important when the word is partially obscured. 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 is significantly harder when they are obscured, at least for humans.

The paper has also highlighted the fact that using context in ML and SR could cause longer learning or processing times for ML and SR systems. However, there is a strong argument for the use of context in accuracy dependent applications, or in those affected by high noise. It is also expected that these issues will reduce, 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 learning and processing.

X. Further Work

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

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

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

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

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

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

[9] 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)*, pp. 98–105, IEEE, 2014.

[10] 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)*, pp. 35–56, 2017.

[11] 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)*, pp. 5020–5024, IEEE, 2017.

[12] 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.

[13] 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)*, pp. 845–848, IEEE, 2002.

[14] 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)*, pp. 24–29, IEEE, 2011.

[15] R. M. Warren and R. P. Warren, "Auditory illusions and confusions," *Scientific American*, vol. 223, no. 6, pp. 30–37, 1970.