2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART) | 978-1-6654-8734-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/SMART55829.2022.10047048

Speech Difficulties and Clarification: A Systematic Review

Ashutosh Dixit[1], Preeti Sethi[2], Puneet Garg[3], Manoj[4] and Juhi Pruthi[5]

[1]Professor, [2]Associate Professor, [3]Assistant Professor, [4]M.Tech. Scholar, [5]Research Scholar

[1,2,4,5]Department of Computer Engineering, J. C. Bose University of Science and Technology, YMCA, Faridabad, Haryana, India

[3]Department of Computer Science & Engineering, ABES Engineering College, Ghaziabad, Uttar Pradesh, India

Email: [1]dixit_ashutosh@rediffmail.com, [2]preetisethi22@gmail.com, [3]puneetgarg.er@gmail.com, [4]mk2757131@gmail.com, [5]juhi.pruthi@gmail.com

***Abstract*—Natural language processing helps the computer to process human languages. In general, human language is presented either in the form of text or speech utterances. Speech based computer interaction system helps the computer to understand human language and perform various tasks for verbal instructions and to make intelligent assistive system. Background noise is the most common factor that causes degradation of the quality and intelligibility of speech. The term background noise refers to any unwanted signal that is added to the desired speech signal. In fact, language description is the main motivation behind widely used machine learning (ML) techniques such as Hidden Markov Models, Discriminant Learning, Structured Sequential Learning, Bayesian Learning, and Adaptive Learning. In addition, speech clarification is being used by machine learning as a large-scale, realistic application to assess a particular method's performance and to address new challenges arising due to the naturally sequential and dynamic character of speech. At the other side, although speech clarification occurs for some applications, it remains largely an unexplained issue for most of the applications. This overview article provides readers an insight about modern ML techniques as they are currently used and relevant to future speech explanation systems. Articles are organized according to the most important ML paradigms that are already popular and can contribute significantly to voice recognition.**

Keywords: Machine Learning, Speech Clarification, De-noising, Natural Language Processing

I. Introduction

Clarification is a fundamental part of man-machine communication and can be available in different forms, as it can have various causes in different degrees of correspondence. It is very clear that explanation is much more important in man- machine communication in comparison to interpersonal communication. This is on the grounds that speech recognition is erroneously skewed and leads to many recognition errors, especially with remote speech. In addition, frameworks may not understand the semantics and setting as people do and as a rule require in general world information. Thus, the dialogue approach is liable for the clarification of uncertain or partial data provided by the customer. Applications such as voice-based calculator need clarification mechanism to understand the user input and perform operation accordingly [1]. A clarification dialogue system, currently used in AI (Artificial Intelligence), once arranged as a design, can turn into an essentially important type of mechanism for managing accidental errors and a variety of issues that typically arise in arguments [2]. Similar to the conventional error of multiple queries, clarification dialogues can make a major contribution to addressing misrepresentation that emerges from the presentation of the questions. The most optimal solution is usually to question the real question maker, for example, to explain whether an asserting claim should be a presumption of the question. In any case, the question clarification may prompt its restructuring, as well as the solution to the issue. Machine learning techniques have made significant contribution in several domains ranging from natural language processing (NLP), computer vision, biomedical, fuzzy logic to intelligent networks [3-5]. This paper addresses the machine learning paradigms adopted by researchers for speech translation and clarification in speech recognition systems.

A. Dialogue System Components

A generic dialogue management architecture makes use of dialogue algorithms generated within the language and domain-free dialogue controller ARIADNE which is explicitly developed for quick prototyping of spoken dialogue frameworks. Dialogue Manager uses Type Feature Structure (TFS) represents semantic input and dialogue information based on typed feature formations. Context-free grammar is employed to parse the user's pronunciation [6]. The language structure is upgraded by data from the ontology characterizing every object, assignment and characteristic that the user can talk about. The parse tree is translated into a semantic characterization following parsing and included in the present dialogue. For speech recognition, a recognition toolkit with the statistical n-gram language models (LM) or Ibis single pass-decoder is used with context independent grammars. Dialogue managers produce these context independent grammars by using the same grammar to transform the resultant parse tree into typed feature formations [7]. the dialogue. The dialogue framework interprets spoken (or typed) input with semantic grammars. Figure 1 illustrates the inclusion of the grammars into the dialogue framework.

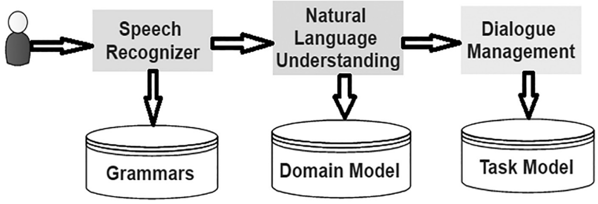

Fig. 1: Inclusion of NLU Elements into the Dialogue Framework

B. *Clarification in Dialogue Sequence*

Clarification dialogues appears to be most firmly connected with the information-oriented dialogue types. The purpose of explanatory dialogue is to assist one side in the conversation by explaining the other side's vague or other challenging statement. To infer such a dialogue and its goal, it has to be understood that there are two sides involved. One of these sides, known as the *respondent,* wants or needs some clarification, while another side, known as *proponent,* is considered to be in a situation to give clarification to the respondent. It is crucial to distinguish the general objective of the dialogue and the separate objectives of each of the two partakers in the dialogue [8]. The overall objective of a dialogue is to effectively complete this process of clarification between the two sides. The proponent gives clarification to the respondent, so as to remove the respondent's confusion or inability to figure out something. The respondent's goal is to get such clarification. Since respondent has no clarity about the speech act, he raises a question, and demands for clarification from the proponent. In order to satisfy this request, the proponent must give some clarification. The respondent will then typically answer by showing whether he felt that the reaction succeeded in addressing his inquiry for clarification or not. Both explanation and clarification include transfer of understanding starting with one party then onto the next in a dialogue.

C. *Need for Speech Clarification*

There are multiple reasons for the need of clarification in interpersonal interaction. Some of the reasons include vague data or acoustical misunderstanding. Communication between man and machine fails even more frequently than interpersonal communication. One justification behind this is the way that speech recognition is flawed particularly in case of distant speech, the input channel is not as clear as for close conversation. Numerous environmental noises show up, brought about by resonance, poor SNR (signal-to-noise ratio), cross talking or other environmental noises [9]. These conditions hinder computerized speech recognition. Hence there is a need to formulate a framework that clarifies the spoken sentence by overpowering the external noise and has network of nodes that store the information until received successfully by the receiver [10-11]. Child's speech clarity in situations like being lost in crowded public place makes them nervous and stammer. In such scenarios, it is essential to have a mechanism that autocorrects the child's spoken speech and words [12]. Speech explanation approaches extract dialogue as the basis for decisions for the subsequent step. At the point of dialogue, the system accesses several dialogue contexts, each characterized by explicit tasks of variables in the extract dialogue condition. The variables' value defines features of the ongoing communication state. The analysis of current and old states helps detecting discrepancies that point out the necessity for clarification [13]. The dialogue situation is properly written as d=(v_1, v_2, v_3, ……., v_n) where each vi is one variable. The abstract dialogue situation employed in the clarification approach comprises of the following variables:

- v_1: INTENTION depicts how well the dialogue data addresses the goal of the client. It is determined based on the conditions of discourse objectives.
- v_2: SELECTED GOALS refers to a set encompassing all objectives that have the state decided. This implies that the discussion suites to these objectives[14].
- v_3: FINALIZED GOALS denotes a set consisting of one or none finished objectives. This intends that there is one objective with state settled and all the data required for execution is available.

II. Literature Review

This section presents an overview of approaches automated as well as generalized adopted by researchers for speech clarification in the literature.

A. *Speech Clarification using Deep Learning*

H. Kuo *et al.* (2014) discussed that the OOV (out-of- vocabulary word) model was enhanced for a S2S (speech-to-speech) translation system in order to play back the audio to the user so that the speech was clarified and corrected [15]. This detector had diverse phases. Initially, a strategy was deployed to recognize a rough location of the OOV. Subsequently, the adjacent decoded words were integrated for covering the true OOV word. A novel CNN (Convolutional Neural Network) algorithm was adopted in real time to illustrate the potential of the presented model. The presented model was applicable to discover diverse parameters which were utilized in diverse phases for clarifying the speech. G. Sterpu *et al.* (2022) constructed a FDNN (fully differentiable neural network) system recognized as Taris to decode the audio-only and audio-visual speech in real time [16]. This system was put together with the traditional algorithm AV for integrating audio-visual speech and recognizing the online speech. An enormous sized public dataset named LRS2 was employed for quantifying the constructed system. The outcomes indicated that the developed integrated system performed more effectively in comparison with others while recognizing the speech for clarifications. Moreover, an AVSR (Audio-Visual Speech Recognition) was deployed for tackling the audio modality in less optimal listening conditions. Z. Eberhart *et al.* (2021) analysed that the dialogue management was employed to verify the process in which a system sent response to user input, such as the possibility of asking a clarification question or displaying the possible outcomes [17]. Thus, a dialogue manager was recommended for interactive API search which concentrated on the search results and dialogue history for selecting the effectual actions. This approach aimed to exploit two policies viz. hand-crafted and a policy whose optimization was done reinforcement learning. M. Korpusik *et al.* (2019) suggested DRL(Deep Reinforcement Learning) to ask follow-up questions in case of recording of a meal description by a user [18]. A new CNN (Convolutional Neural Network) algorithm was implemented for avoiding the typical FE (feature engineering) which assisted the dialogue systems in handling the text mismatch amid the NL(natural language) user queries and structured database entries. Moreover, a RL(reinforcement learning) agent was generated for following up with the user. The results depicted that the suggested approach led to enhanced recall value around 89.0% in the speech clarification. In addition, the hybrid RL technique offered higher naturalness ratings in a human evaluation. Table-1 summarizes the deep learning methods used by researchers for speech clarification.

Table 1: Speech Clarification using Deep Learning Techniques

Author	**Technique Employed**	**Evaluation Parameters**
H. Kuo *et al.*	OOV (out-of-vocabulary word) model and CNN (Convolutional Neural Network)	Jaccard metric (95%), recall (78%) and FPR (False Positive Rate) (6%)
G. Sterp u, *et al.*	FDNN (fully differentiable neural network) system	Accuracy (98%), recall (86.3%)
Z. Eberhart *et al.*	Reinforcement learning	Precision (90.5%) and accuracy (89%)
M. Korpusik *et al.*	DRL (deep reinforcement learning)	Recall (89.0%)

B. Speech Clarification using Automatic Speech Recognition System

N. F. Ayan *et al.* (2013) introduced a new technique in order to enhance the communication success among users of S2S (speech-to-speech) translation systems for which the errors were detected in the output of ASR (automatic speech recognition) and SMT (statistical machine translation) algorithms [19]. The system-driven targeted clarification was started for clarifying the erroneous areas in user input and repairing them based on the given user responses. The unbiased subjects were employed in live mode for computing the introduced technique. The results validated that the introduced technique was useful to enhance the efficacy of communication among users of the system. Moreover, this technique offered accurate transcription in 73% of the test sentences after accomplishing the clarification attempt. S. Stoyanchev *et al.* (2015) established a new method to deal with the errors obtained in ASR and NLU process [20]. For this, TC(targeted clarification) was deployed in an interactive spoken dialog system. The clarification question raised on the misrecognized portion of the utterance was considered to implement TC in case the spoken utterance was found partial. The key element in this process was an accurately detected value of localized ASR and NLU errors in a speech. An interactive multimodal assistant was employed in the quantification of the established models. After detecting the presence and correctness based on Oracle, the established method with a TC became capable of clarifying 38% of errors. T. Shinozaki *et al.* (2016) projected an enhanced CRECA (context respectful counselling agent) to extract emotional words from the speech of clients throughout their dialogue for detecting their changes and offering changes to the clients as dialogue summary [21]. For continuing the conversation, the projected technique was effective for constructing or relating the topics with eventual and emotional words in dialogue sentences. Thereafter, a fuzzy reasoning was presented for selecting a non-boring response various digging prompt or from numerous sentences in case the input utterance was not matched with the earlier patterns. Hence, the projected technique offered clarification and self-awareness which helped in dealing with more complex issues. Y. Yamamoto *et al.* (2015) designed an enhanced CRECA (context respectful counselling agent) which was effective for extracting the emotional words from the speech of users during their dialogue [22]. Hence, the changes were detected. In case of failure of detecting the changes, this algorithm sent response using paraphrases of clients. This algorithm was able to pretend as it recognized the mental issues of clients as contexts or situations. Hence, the clients got capacity of performing clarification and self- awareness that assisted in addressing both the issues. Ashutosh *et al.* propose an application *Rakshak* that accepts the vocal information provided by lost children, clarifies the misspelt or unclear words and helps in decoding the relevant details [10]. The speech-to-text output helps the police officials in preparing the search corpus and reuniting the lost children with their parents.

C. Speech Clarification using General Techniques

A. Antenucci *et al.* (2021) investigated an advanced robotic security solution planned on the basis of IoRT

(Internet of Robotic Things) paradigm [23]. This approach presented the robotic guards relied on VIKI (Vitrociset AI) model with an integrated biometric module for generating ePassport instant images, that were downloaded from RFID. A human interacting module containing the speech synthesis functions was employed to carry out QA (question- answering) sessions during controlling the access for interactive interviews. This approach helped individual in asking questions to the robotic guards, for interacting and collaborating with humans. For this, the security questions and clarifications were inserted while providing access control to restricted and endangered regions. S. Amiri *et al.* (2019) presented a dialog agent for robots for interpreting user commands with the help of a semantic parser, when a probabilistic dialog manager was employed to ask the clarification questions [24]. This agent was applicable for augmenting its knowledge base and enhancing its language capabilities. MTurk and real-robot platforms were applied to conduct simulations for computing the presented system. The simulation results exhibited that the presented system was more effectual and accurate in comparison with other methods. T. Shao *et al.* (2021) discussed that the system concentrated on automatically generating the clarification questions so that misunderstanding was avoided. Thus, a SHiP (Self-supervised Hierarchical Pointer-generator) method was suggested [25]. Similar to CoF (Coarse-to-fine) procedure of CQG(Clarification Question Generation), two operations such as to predict the dialogue history and predict the EN (Entity Name) were formulated. Afterward, a HT(Hierarchical Transformer) model was integrated with a PG (pointer-generator) with the objective of understanding the ambiguous multi-turn conversations and addressing the issue of OOV(out- of-vocabulary word). In the end, an E2E (end-to-end) co-training paradigm was suggested for training the pretext and downstream tasks. The results acquired on CLAQUA depicted that the suggested approach enhanced the BLEU (BiLingual Evaluation Understudy) up to 6.75% and ROUGE-L(Recall- Oriented Understudy for Gisting Evaluation) by 3.91%. T. Shinozaki *et al.* (2016) introduced an enhanced CRECA (Context Respectful Counselling Agent) for accomplishing an objective of lingualized in conversation [26]. The psychological issues were asked in this approach that offered a wide perception. This process assisted in mitigating the time to perform the problem-solving task. The introduced approach offered speedy alertness and alleviated the mental and physical burden of client. General techniques used in speech clarification are summarized in table-2.

TABLE 2: SPEECH CLARIFICATION USING GENERAL TECHNIQUES

Author	**Technique Employed**	**Evaluation Parameters**
A.Antenucci *et al.*	Advanced robotic security solution	Accuracy (93.4%)
S. Amiri et al.	A dialog agent	Confidence level (0.1), F1 score (0.79)
T. Shao *et al.*	SHiP (Self- supervised Hierarchical Pointer -generator) method	BLEU (6.75 %) and ROUGE-L (3.91 %)
T.Shinozaki *et al.*	An enhanced CRECA (Context Respectful Counselling Agent)	F1 score (0.62), precision (98%)

III. CONCLUSION

Although speech intelligibility can be thought of as a component of quality, high-quality speech always has good intelligibility. Therefore, no speech augmentation tool can enhance both speech quality and understandability. Clarification dialogues are an effective and straightforward way of dealing with speech recognition errors and can be applied to a variety of speech interface applications. Over the past decades, many techniques have been used to improve speech quality and understandability, with time-domain and frequency-domain algorithms being the two important categories. Novel insights from contemporary ML methodology hold great potential to advance the cutting edge in speech clarification technology. A speech clarification problem can be viewed as an application of ML, just like computer vision, bioinformatics, and NLP (Natural Language Processing). When viewed in this way, speech recognition and clarification is a particularly useful ML application because it has a large training and testing corpus, is computationally intensive, has a unique sequential structure in the inputs, and is characterized by structured outputs. It is analysed that machine learning techniques are the most efficient techniques for the speech clarification. ML based approach addressed the emotional and mental states of the human subjects that aid in resolving their mental issues and enhancing wellness.

References

[1] P. Sethi, P. Garg, A. Dixit and Y. Singh, (2020, April). Smart number cruncher–a voice based calculator. In *IOP Conference Series: Materials Science and Engineering* (Vol. 804, No. 1, p. 012041). IOP Publishing.

[2] E. Knauss, D. Damian, G. Poo-Caamaño and J. Cleland-Huang, "Detecting and classifying patterns of requirements clarifications," 2012 20th IEEE International Requirements Engineering Conference (RE), 2012, pp. 251-260.

[3] P. Garg, A. Dixit and P. Sethi, (2022). ML-Fresh: Novel Routing Protocol in Opportunistic Networks Using Machine Learning. *Computer Systems Science & Engineering, Forthcoming.*

[4] P. Garg, A. Dixit and P. Sethi (2021, May). Link Prediction Techniques for Opportunistic Networks using Machine Learning. In *Proceedings of the International Conference on Innovative Computing & Communication (ICICC).*

[5] K.A. Shastry and H.A. Sanjay, (2020). Machine learning for bioinformatics. In *Statistical modelling and machine learning principles for bioinformatics techniques, tools, and applications (pp. 25-39). Springer, Singapore.*

[6] E. Knauss and D. Damian, "V:Issue:lizer: Exploring requirements clarification in online communication over time," 2013 35th International Conference on Software Engineering (ICSE), 2013, pp. 1327-1330.

[7] Lei Guo, Xiaodong Wang and Jun Fang, "Ontology Clarification by Using Semantic Disambiguation," 2008 12th International Conference on Computer Supported Cooperative Work in Design, 2008, pp. 476-481.

[8] C. Lewis and G. Di Fabbrizio, "A clarification algorithm for spoken dialogue systems," Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., 2005, pp. I/37-I/40 Vol. 1.

[9] M. Kulkarni and S. Sane, "An ontology clarification tool for word sense disambiguation," 2011 3rd International Conference on Electronics Computer Technology, 2011, pp. 292-296.

[10] P. Garg, A. Dixit and P. Sethi (2021, April). Opportunistic networks: Protocols, applications & simulation trends. In *Proceedings of the International Conference on Innovative Computing & Communication (ICICC).*

[11] P. Garg, A. Dixit and P. Sethi (2019). Wireless sensor networks: an insight review. *International Journal of Advanced Science and Technology, 28*(15), 612-627.

[12] A. Dixit, P. Sethi and P. Garg (2022). Rakshak: A Child Identification Software for Recognizing Missing Children Using Machine Learning-Based Speech Clarification. *International Journal of Knowledge-Based Organizations (IJKBO), 12*(3), 1-15, Jul 2022.

[13] G. M. Kruijff, M. Brenner and N. Hawes, "Continual planning for cross-modal situated clarification in human-robot interaction," RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication, 2008, pp. 592-597.

[14] S. Stoyanchev, P. Salletmayr, J. Yang and J. Hirschberg, "Localized detection of speech recognition errors," 2012 IEEE Spoken Language Technology Workshop (SLT), 2012, pp. 25-30.

[15] H. Kuo, E. E. Kislal, L. Mangu, H. Soltau and T. Beran, "Out-of-vocabulary word detection in a speech-to-speech translation system," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 7108-7112.

[16] G. Sterpu and N. Harte, "Taris: An online speech recognition framework with sequence-to-sequence neural networks for both audio-only and audio-visual speech", Computer Speech &Language, vol. 12, no. 7, pp. 1572-1579, 19 January 2022.

[17] Z. Eberhart and C. McMillan, "Dialogue Management for Interactive API Search," 2021 IEEE International Conference on Software Maintenance and Evolution (ICSME), 2021, pp. 274-285.

[18] M. Korpusik and J. Glass, "Deep Learning for Database Mapping and Asking Clarification Questions in Dialogue Systems," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 8, pp. 1321-1334, Aug. 2019.

[19] N. F. Ayan *et al.*, ""Can you give me another word for hyperbaric?": Improving speech translation using targeted clarification questions," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 8391-8395.

[20] S. Stoyanchev and M. Johnston, "Localized error detection for targeted clarification in a virtual assistant," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5241-5245.

[21] T. Shinozaki, Y. Yamamoto, S. Tsuruta, E. Damiani and R. Knauf, "Highly enhanced context respectful counselling agent," 2016 IEEE International Conference on Fuzzy Systems (FUZZIEEE), 2016, pp. 2174-2181.

[22] Y. Yamamoto, T. Shinozaki, S. Tsuruta, E. Damiani and R. Knauf, "Enhanced Context Respectful Counselling Agent," 2015 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), 2015, pp. 246-253.

[23] A. Antenucci *et al.*, "An Industrial Distributed Network of Intelligent Robotic Security Guards Based on the Internet of Robotic Things Paradigm," 2021 International Conference on Computer, Control and Robotics (ICCCR), 2021, pp. 9-13.

[24] S. Amiri, S. Bajracharya, C. Goktolgal, J. Thomason and S. Zhang, "Augmenting Knowledge through Statistical, Goal-oriented Human-Robot Dialog," 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019, pp. 744-750.

[25] T. Shao, F. Cai and H. Chen, "Self-supervised clarification question generation for ambiguous multi-turn conversation", Information Sciences, vol. 12, no. 7, pp. 672-680, Dec. 2021.

[26] T. Shinozaki, Y. Yamamoto, S. Tsuruta, Y. Sakurai, E. Damiani and R. Knauf, "Goal Aware Context Respectful Counseling Agent," 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), 2016, pp. 252-257.