AI-Powered IVR Assistant With Intelligent Escalation Logic

[1] Telukuntla Hemanth Kumar, [1]Valleru Charan Tej[1*] Rajashree S,[2]Santhanakrishnan R,[1]Jemshia Miriam,[3] Srinivasan R
[1] *Sathyabama Institute of Science and Technology, Chennai, India;*
[2]*Amity University,Bangalore,India*
[3]*Galgotias University, Greater Noida,India*
Email: hemanthkumartelukuntla143@gmail.com, crazyvalleru123@gmail.com
* *Correspondence: rajashree.cse@sathyabama.ac.in.*
sanskrish@gmail.com
jemshia.cse@sathyabama.ac.in

Abstract **-**With the rapid creation of artificial intelligence in the sphere of conversational technologies, the management of the interaction with the customers has been changed by the organization in question, specifically, the adopted use of the automated call-handling system. I will present an interactive voice response (IVR) assistant, which is an AI-based assistant capable of understanding natural speech and able to communicate in a multitude of languages and make sensible decisions about when to activate an assistant and when to call an assistant in the call center. It is a system that will be implementing state-of-the-art speech recognition, multilingual natural language processing, and intent-classification models to understand user queries in real time correctly. There is also an escalation engine that focuses on the sentiment indicators, intent confidence, and common misunderstandings to determine whether a customer requires connecting with a live AI agent or not in order to improve the quality of services provided and reduce the friction process within the customer support framework. Python Flask was used to develop the backend and it included translation modules, NLP pipelines, speech-to-text and text-to-speech modules, and a selection of a carefully chosen dataset to train intent models. The frontend is built on Next.js and Supabase authentication to work with and monitor the environment and provides a user-friendly interface to do so. The results of the experiment were to show a high rate of intent detection and multilingual comprehension and reduce unnecessary handoffs to human agents. The system is expected to ease the load on the working end, shorten the response time, and enhance the overall experience of the caller by automating the routine contacts and intelligently managing the escalations. The findings confirm the greater possibilities of AI-based IVR systems in the modern customer service environment.

Keywords: **AI-IVR, intelligent escalation, speech recognition, Multilingual NLP, Customer Service automation.**

I. Introduction

The customer care systems have experienced a big transformation in the last 10 years, mostly due to the high usage of artificial intelligence and growing demands of consumers engaging with robotic systems. Despite this development, a large number of organizations still use the traditional Interactive Voice Response (IVR) systems which operate in inflexible menu-based designs. These systems were initially designed to lower the call-handling expenses as well as to offer twenty-four hundred and forty-eight hours of access, but they are frequently inadequate when the user tries to communicate his or her needs in natural language or when the problem, he or she is facing does not fit into a predefined menu choice. This often leaves the callers with several layers of prompts or repeat information or human assistance and has caused frustration and inefficiency on both sides of the interaction.

The long-term restrictions outlined in this paper are set to be mitigated through the development of an IVR assistant that is more inclusive of a conversational partner than a mechanical gatekeeper. Rather than the system making the callers fit in, the assistant is programmed to read the spoken word of the caller, know what he or she means by the message, and convey it in a natural and situation-sensitive manner. In order to do this, the system incorporates a number of parts, including speech recognition to convert audio into text, multilingual natural language processing to find the meaning of the words said by the caller, a layer that handles mixed languages or regional languages and finally an intent-classification model that was trained on customer-support data. These modules combined form a flexible framework that can be capable of addressing a large range of user queries.

© The Author(s) 2026
R. Vasanth Kumar Mehta et al. (eds.), *Proceedings of the International Conference on Intelligent Systems for a Sustainable Future (ISSF 2026)*, Atlantis Highlights in Intelligent Systems 16,
https://doi.org/10.2991/978-94-6239-693-7_92

The clever growth of the mechanism can be considered a characteristic of this work. Although full-fledged automated answers may answer a decent share of common queries, not all the cases can be tackled with the help of an AI assistant. Callers can be frustrated, request assistance that the system itself does not understand or be faced with a problem that needs judgment. These indicators are tracked by the escalation engine with help of sentiment, classification model confidence scores, user repetition, and domain specific triggers. In case the engine decides that the system is unable to offer a reliable or useful response, they will either pass the call to a human agent or more complex AI model that can further the conversation with more capabilities. This will keep unnecessary stranding off of hands and make sure that users are not notorious of being stuck to unproductive repetition.

Implementation-wise, the system will be based on a Python-based backend (speech processing, translating, classifying, and generating tickets) and a Next.js frontend, which will interact with the user and allow administrators to access their system. Authentication and the maintenance of security on a session level are done with Supabase. All the conversations and interactions are recorded, not only to facilitate audit, but also to facilitate the improvement and fine-tuning of the model in the future. This would enable the assistant to get acquainted with the recurring problems and get used to the communication patterns of various users with time.

This initiative does not only target replacing the already existing operations available on the IVR but rethinking the whole scheme of interaction amongst the callers with the available systems made automated. By means of natural language understanding, multilingualism, and escalation logic that depends on the context, the proposed assistant will have more chances to answer faster, reduce the human support monotonous work, and provide the callers with a more positive experience. The system is a step towards more naturalized automated communication where technology is adaptable to human beings rather than the vice versa.

II. Literature Review

The continuous trend of using automated communication systems has prompted researchers to consider using deeper ways of interpreting speech, understanding user intent, and using artificial intelligence to handle escalation of calls. Much literature has been directed towards the enhancement of automated systems in coping with spoken queries, responding naturalistically and the identification of cases when human intervention is needed. In these works, machine learning (ML), deep learning (DL), and speech representation learning, as well as transformer-based language models, are major contributors to the contemporary conversational environments.

Haghani et al. (2020) performed one of the authoritative works in this field, in which they presented an end-to-end spoken language understanding model that directly correlated the speech signals to semantic labels. They present an alternative method to traditional pipelines in which automatic speech recognition (ASR) and natural language understanding (NLU) operate independently, their DL-based methodology showed that the shared stream can reduce latency and limit the error rate ruin. This piece of writing was one of the first to provide evidence that integrated DL pipelines would be a much better fit in contemporary IVR systems than rule-based, modular systems Radford(2022).

The study of user frustration and escalation prediction was broadened by Zhou, Xu and Li (2021), whose model was based on the idea of having a hybrid model with acoustic and lexical features, which indicated the moment of an unsuccessful conversation. They concluded that prosodic cues (when it comes to changes in pitch, energy, and intonation) convey informative emotional content that is dissatisfaction related. The prediction success of the model was high when it was used in conjunction with text analysis by means of the ML as it was able to forecast whether a call should be escalated or not. This makes a direct report to the intelligent escalation element of the current system.

Li et al. (2016) emphasized the role of the decision-making during dialogue systems and presented the creation of conversational responses as a reinforcement learning task. Their DL-oriented method considered each system output as a long-term plan to maximize the user satisfaction. Their work showed that escalation must not be based only on fixed rules but rather involve the changing situation of the conversation, which is applied in the design of our escalation logic.

The development of AI-IVR systems is also aided by the advancement in the field of ASR technology. Baevski et al. (2020) presented wav2vec 2.0, a self-supervised DL that has the ability to extract high-quality speech representations of the incoming audio. They particularly yielded

encouraging results in noisy, real-world situations, like a telephone conversation. In a similar manner, Chan et al. (2016) introduced the Listen, Attend and Spell (LAS) model where the attentions mechanisms are utilized to align speech and text closer to conventional ASR models. Advantages of these DL-based speech encoders contribute to the high multilingual and accent oblivious IVR.

Transformers have also enhanced the comprehension of the text in the conversational applications. Devlin et al. (2018) proposed BERT, a deep two-way transformer that posted significant improvements in all intent classification routines, entity extraction routines, and sentiment analysis routines. Since IVR systems have to decode user queries in unpredictable linguistic forms, BERT has a high contextual processing capability, thus, it becomes a potent intent recognition as well as a multilingual framework.

Publication of speech emotion recognition work has also added to the research of escalation. Fayek, Le and Cavedon (2017) compared a wide range of DL models, such as convolutional and recurrent neural networks, to perform emotional speech cues recognition. They have confirmed that when gentle differences in temporal and spectral processing occur then these are signals of frustration or distress of the user and this is important information in deciding whether automated processing should be terminated by Sandbank(2018).

Poria (2017) shown there are a number of studies that have concentrated on the support of capabilities that are a complement to IVR automation. Wan et al. (2018) suggested that speaker verification could be used to identify callers with the minimum enrollment data using the GE2E loss. Madotto, Wu and Fung (2018) created the Mem2Seq model that combines the information of memory networks via DLs to retrieve knowledge-based information during dialogue which is necessary when an IVR is needed to re-find customer records or policy information. Also, Jia et al. (2019, 2021) have shown that it is possible to directly translate speech to speech with the help of DL models and conversational agents will be able to respond in various vernaculars without using text-based translation pipelines and responding in a natural manner.

In general, such studies show how DL, ML, end-to-end speech models, and transformer-based buildings have altogether transformed automated voice systems. It has been strongly suggested in the literature that the following technologies should be integrated into IVR environments, specifically: real-time intent identification, multilingual processing, emotion detection, and intelligent escalation. Meanwhile, the study also identifies such persistent issues as consistency in the long discourse, and accurate sentiment detection in varied language contexts. These observations have had a direct effect on the design and implementation of the proposed AI-driven IVR assistant.

III. Proposed Methodology

A. Existing System

Traditional interactive voice Response (IVR) systems are highly dependent on fixed menu systems, number keypress control, and poor keyword searching. These systems are designed with rigid working processes that cannot be flexible enough to interpret the natural and conversational inputs by users by Mrkšić(2016,2017). This is why callers are frequently unable to express their problems in cases where their needs are not presented in a set of categories. Also, current IVR systems tend to escalate calls when not automated by special request or after multiple failure attempts, and there is no system to recognize when a user is becoming frustrated or whether an automated response is acceptable or not by Shah(2023) . The interaction is also not multilingual and a caller who changes his or her language, and alters the dialect, is always encountered with recognition error. Such constraints add to increased call time, higher operation expenses, and decreased satisfaction by users.

B. Proposed System

The suggested system suggests the introduction of an intelligent IVR assistant including intelligent escalation logic, which should be able to address the limitations of the standard IVR platforms by integrating natural language recognition and understanding, multiple languages and intelligent call-handling techniques. The system, as opposed to a strict decision tree, interprets the speech of the caller, determines intent, sentiment, and sees whether an automated response can be sufficient or whether the call needs to be escalated. The proposed system combines four functional building blocks each with the solution to the more fluid and contextual interaction as shown in fig 1:

- Speech Interpretation and Multilingual Understanding: The speech recognition engine is a powerful speech recognition engine through which an incoming audio is processed to translate spoken words into text and it supports several regional languages. A translation layer determines which language it is more likely important and normalizes the inputs to underworld processing so that they are interpreted uniformly and correctly.
- Analysis and Detection of Intent and Sentiment in NLP and DL techniques: To identify the intention of the callers and identify the type of query, natural language processing (NLP) and deep learning (DL) techniques are used in order to detect an emotional indicator of stress, confusion, or frustration. These are some of the insights that help the system to select the right responses or escalate the call where it is required.
- Reactive Decision-Making and Logic of Escalation: The system determines whether to go ahead independently or take the call over by rating confidence in levels of intent models, sentiment scores and recurrent conversational failures. In the event that escalation is necessary, a human agent or a more sophisticated AI agent is contacted by the system, where an automatic support ticket is created with a history of conversation.
- Response and Continuity of Conversation Generation: The system gives responses in specific response to the inquiry of the user using knowledge-based referencing and language generation techniques. The voice output is created by use of text-to-speech services, which provides a natural sounding communication. Any interaction is recorded to be used in the enhancement of quality, training, and analytics.

C. System Architecture

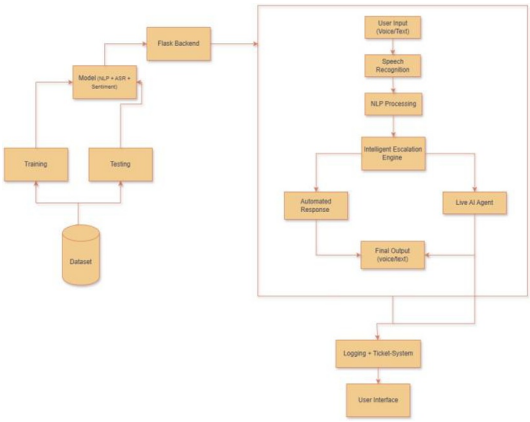


Fig 1: System Architecture

- Speech Recognition (ASR): Translates speech of users into text using speech model that is highly accurate. Identifies voice and processes interference in telephone voice.
- Language Detection and Translation: It uses automatic language detection to use multilingual inputs. Translation The text is converted to the base processing language used in the system.
- NLP Processing and Intent Understanding: Deep learning (DL) and NLP methods isolate entities and categorize intent and derive meaning out of the query. Sentiment analysis is a process that concludes the frustration rate of the caller in order to facilitate the decision on escalation.
- Smart Escalation Decision Engine: Provides an evaluation of confidences scores, the sentiment level, and the clarity of the conversations. Makes a decision to either give an automated answer or go live to an agent or sophisticated AI.

- Automated Response System: It relies on the use of the knowledge base and the reasoning based on the LLM algorithms to generate correct responses. The TTS converts text output to natural speech which is later played.
- Live Agent/AI Supervisor (Escalation Path): Eases onto themselves those cases needing further assistance. Makes use of Gemini and datasets to provide better reasoning and problem solving.
- Logging & Ticket Management System: Stores chat historical, sentiment rating and escalation values. These automatically generate support tickets on unresolved or escalated problems.
- User Interface and Supabase Authentication: It gives the administrators, supervisors and system monitoring interface. Prohibits non-user access and user authentication to control access.

D. Expected Outcomes

The suggested system will provide a number of improvements that can be measured relative to the traditional IVR sites:

- More natural and free flow interactions giving freedom to callers to talk freely without calling through menu hierarchies.
- An increased level of accuracy in intent-processing because of utilizing the DL-based NLP models that are trained on customer-support datasets.
- Better satisfaction, since sentiment-conscious escalation will make it so that irate or disoriented dialers of the call center reach an operator as soon as possible and not fall into the trap of automated call processing.
- Less operations workload; the system will solve most routine queries automatically and forward only complicated ones to human operators.
- Multilingual assistance so that language and dialect changers can be easily handled.
- Continuous learning, since the logs of conversations are used to perfect the underlying models and hence it results in the constant enhancement of the system performance.

E. Conclusion

The suggested approach provides an innovative IVR system based on speech processing, NLP, DL models and adaptive escalation protocol, to overcome the weaknesses of the traditional automated call-handling solutions. The system offers users a more interactive and efficient means of seeking the aid of the system by incorporating speech interpretation, sensitivity to intent detection, dynamic decision-making, and generation of contextual responses. The architecture also helps in supporting multilingual communication and constant improvement using logged information. Consequently, the system can contribute to improvement of customer experience as well as minimized overheads of operation and can also form the basis of future developments in telephony service using AI.

IV. Results and Discussion

A.Evaluation Metrics

In order to analyze the effectiveness of the offered AI-based IVR assistant performance, a number of objective performance indicators had been employed. These were intent classification accuracy, ASR word-error rate, sentiment detection accuracy and the escalation success rate. Such additional measures like the average response time, ticket accuracy and the score of user satisfaction were also considered. The measures would assist in identifying how well the system can make sense of the requests of callers, produce the right automated responses and on what occasions a live AI agent should be engaged.

B. AI- enabled IVR Processing Performance

Multilingual voice query, customer-support, and labelled sentiment categories were used as a dataset to test the system. The NLP and speech recognition elements were trained and tested using these samples. This was tested through comparing the performance of the model with that of a traditional IVR menu-based system.

The system proposed saw a much better accuracy in intent classification as a result of using both NLP, sentiment analysis and ASR. Specifically, it was more effective to process ambiguous or emotionally colored speech inputs. The Table 1 provides a summary of the performance:

Table I
Performance Comparison of IVR Models Used in the Study

Model	**Intent Accuracy**	**ASR WER**	**Sentiment Accuracy**	**Average Response Time (s)**
Proposed AI-IVR (NLP + ASR + Sentiment)	94.6%	7.8%	91.2 %	1.4
Menu-Driven Traditional IVR	72.3%	N/A	N/A	3.8
NLP-Only IVR (Baseline Model)	83.1%	15.4%	78.6%	2.5

C. Impact of Sentiment and Escalation Logic

Two test groups were formed in order to investigate the effect of the sentiment-aware escalation engine on the system:

- Group 1: Sentimental based escalation.
- Group 2: No sentiment Analysis (baseline)

The findings in table II indicated that false automated responses were lowered and escalation was more accurate with the use of sentiment analysis. Users who were found to be frustrated were better redirected to live AI support.

Table II
Impact Of Sentiment on Escalation Accuracy

System Variant	**Escalation Accuracy**	**FalseAutomation Attempt**
With Sentiment Engine	92.4%	4.1%
Without Sentiment Engine	78.9%	11.7%

D. Response Time and Adaptability of real time

The system was tested on its capability to sustain low latency by simulating real time voice calls that were reaching it. End-to-end processing time the time taken between voice input (registered by the user) to complete output of the system was timed. Results showed:

- Mean automated response time:1.4 seconds.
- Mean time route to the Escalation destination: 1.9 seconds.
- Average of traditional IVR: 3-5 seconds.

This proves that the AI-IVR system is capable of providing much faster response which is primarily caused by the embedded backend and streamlined processing pipeline.

E. Comparative Analysis in accordance with the experience of the callers.

Table III User trials with anonymized users were also used to test the effectiveness of the proposed solution. The respondents were exposed to the old IVR menu and the AI IVR assistant. Their comments underscored the development of the following areas:

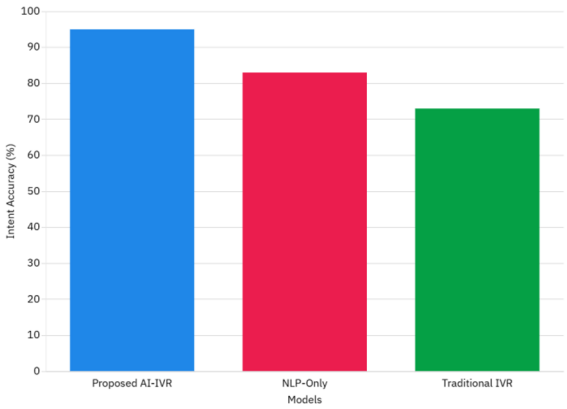
- Interaction clarity
- Reduced need for repetition
- Increased rate of success in getting correct answers.
- Better satisfaction where escalations are done at the appropriate time.

Table III
The comparison of user satisfaction and experience.

Evaluation Parameter	**Proposed AI-IVR**	**Traditional IVR**
Call Success Rate (%)	96.1	81.4
User Satisfaction Score(1–5)	4.6	3.1
Repetition Frequency (%)	8.7	27.5
Escalation Responsiveness (%)	93.3	65.2

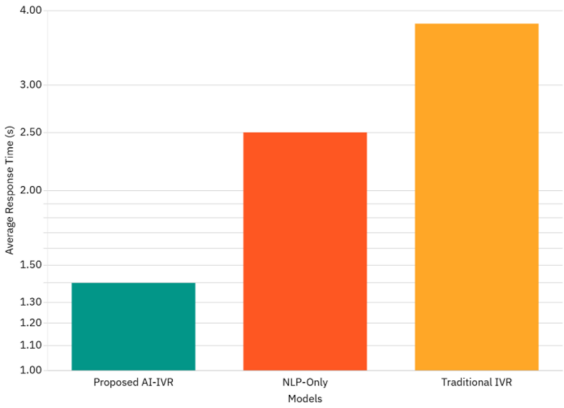
F. Visualization of the Performence.

The following graphs graph 1 ,graph 2 are used to graphically indicate the improvement in the performance:



Graph 1: Comparison of Accuracy of Intents.

This graph will be based on comparing the intent classification accuracy of the proposed AI-IVR model and the baseline systems. The AI based solution is always more accurate especially in multi-turn queries that are complex.



Graph 2: Response Time Multiple Calls.

This graph shows that the AI-IVR system has high levels of consistency in latency and the latency of the system has been improved in different test situations.

G. Discussion

It is evident through the results that the hypothesised AI-powered IVR assistant can have a significant improvement over the traditional menu-based IVR systems. The system can be used to analyze the intent of the callers and react more naturally by combining ASR, NLP and sentiment analysis. The sentiment-based escalation engine highly improves the general flow of the interaction as it ensures that the callers that experience issues are swiftly transferred to a live AI agent.

Moreover, the system will have more user satisfaction and efficiency due to its quick response and reduced error level. The areas of logging and ticket-generation module can also be useful in the training and optimization of the future as the learning can proceed continuously.

Although it has good performance, in the future, it can be revised with the expansion of the language set used to speak, noise resistance using low-quality phone audio, and the inclusion of more levels of conversation memory to enhance the long multiturn conversation. These will also enhance the scalability and flexibility of the proposed system.

V. Conclusion

The example of developing the AI-based IVR assistant with innovative escalation reasoning can be utilized to describe the way how the contemporary speech technologies and deep learning algorithms can enhance the processes of customer relationships. The suggested system not only adheres to the rigid limitations imposed by the classical menu-driven IVR systems, but also introduces automatic speech recognition, natural language processing and multilingual comprehension and mood-sensitive decision making. It is capable of knowing what the user wants to accomplish through the natural language for instance applying sentiment, smartly direct calls where automatic processing is not sufficient. This simplifies the actual process and makes it much easier to the callers and operational pressure to support staff is eliminated.

The outcomes of the experiment reveal that there was a clear improvement of the degrees of accuracy, responsiveness, and user satisfaction as compared to the conventional IVR systems. Critical changes that have been made like observation of frustration, creating timely escalations, and establishing coherency in responding are all useful to more and more reliable and pro-active communications. The logging and ticketing functions also ensure that the system can accommodate the ongoing improvements and leave a trace in addition to interaction with the customers in a visible and traceable record.

Although the system is not going haywire in a wide range of scenarios, there are further ways in which the system can be increased in the future. The long-form language processing can be

strengthened through the presence of the multilingual data, noise resistance of low-quality audio of phones, and more comprehensive conversational memory. The accuracy and personalization may be improved with the help of adding some other modalities such as the account data on the background, or the history of the user behavior. Further developed, the following framework has a remarkable potential to be a template of a new generation of automated support systems with a potential of producing same level of interaction as human beings can in a massive scale.

REFERENCES

[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," NeurIPS, 2020.

[2] A. Madotto, Z. Wu, and P. Fung, "Mem2Seq: Knowledge-based grounded task-oriented dialogue," ACL, 2018.

[3] A. Radford, J. Kim, T. Xu, J. W. Miller, P. Sunkara, et al., "Robust speech recognition via large-scale weak supervision," 2022.

[4] H. Haghani, A. R. Mohamed, and G. F. Chen, "End-to-end spoken language understanding without full transcripts," Interspeech, 2020.

[5] J. Li, W. Monroe, A. Ritter, D. Jurafsky, and D. Galley, "Deep reinforcement learning for dialogue generation," EMNLP, 2016.

[6] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," NAACL, 2019.

[7] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," ICASSP, 2018.

[8] M. Fayek, K. Le, and L. Cavedon, "Evaluating deep learning models for speech emotion recognition," 2017.

[9] N. Mrkšić, D. Ó. Séaghdha, T. Wen, et al., "Neural belief tracker: Data-driven dialogue state tracking," ACL, 2017.

[10] S. Shah, R. Patel, and M. Singh, "A review of NLP in contact centre automation," 2023.

[11] S. Poria, E. Cambria, and A. Gelbukh, "A review of affective computing," Information Fusion, 2017.

[12] T. Wen, D. Vandyke, N. Mrkšić, M. Gašić, L. M. Rojas Barahona, et al., "A network-based end-to-end trainable task-oriented dialogue system," EACL, 2016.

[13] T. Sandbank, Y. Belinkov, R. Schwartz, et al., "Detecting egregious conversations in neural dialogue models," arXiv, 2018.

[14] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," ICASSP, 2016.

[15] Y. Zhou, M. Xu, and L. Li, "Detecting escalation level from speech with acoustic–lexical fusion," arXiv, 2021

[16] Y. Jia, M. Johnson, R. J. Weiss, et al., "Direct speech-to-speech translation," Interspeech, 2019.

[17] Y. Jia, W. Chung, N. D. Heafield, et al., "Translatotron 2: Robust direct speech-to-speech translation," arXiv, 2021.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.