Interspeech 2022
18-22 September 2022, Incheon, Korea

Spoken Dialogue System for Call Centers with Expressive Speech Synthesis

Davis Nicmanis[1], *Askars Salimbajevs*[1,2]

[1]Tilde SIA
Vienibas gatve 75a, Riga, Latvia
[2]Faculty of Computing, University of Latvia
Raina bulvaris 19, Riga, Latvia
{davis.nicmanis, askars.salimbajevs}@tilde.lv

Abstract

In this paper, we present a prototype of a spoken dialogue system that integrates automatic speech recognition (ASR), natural language understanding (NLU), bot management system, and expressive text-to-speech (TTS). Such a solution can be integrated into a call center to provide first-line support, replace older interactive voice response (IVR) systems, decrease the load on call center operators and greatly improve client experience.

The prototype is primarily designed for the Latvian language, however, support for other languages can be easily added by replacing language-specific components like ASR and TTS.

Index Terms: speech recognition, human-computer interaction, expressive speech synthesis, dialogue system

1. Introduction

With the advent of the pandemic, many aspects of our lives became remote. For example, in Latvia, this led to a surge of interest in chatbots - virtual assistants that help users navigate websites of government agencies and private companies. Most such chatbots are text-based dialogue systems, however, for many people (especially, seniors) voice is the preferred way of communication. Therefore, there is a growing interest in Latvia (and worldwide) to integrate such virtual assistants into support call centers, where bots can replace older interactive voice response (IVR) systems, decrease the load on operators and greatly improve client experience.

In our previous work [1], we developed an expressive Latvian speech synthesizer for our text-based dialog system to enable voice feedback that can match the sentiment of the conversation. However, the communication with the bot was still text-based from the user side. This paper presents a prototype of a fully voice-based bot solution that integrates automatic speech recognition, natural language understanding, a bot management system, and expressive text-to-speech.

Because such technology is not available for the Latvian language, the prototype was primarily designed for the Latvian, however, it can be extended with support for other languages.

The presented prototype allows to design dialogues and fine-tune intent detection models using intuitive graphical user interface. Importantly, a dialogue designer can tailor the sentiment of each response using expressive speech synthesis. We believe that expressivity is an essential part of human verbal communication. For example, it can be used to show empathy or convey additional context to the information. In the video, we demonstrate the proposed solution in a call center scenario.

2. Method

2.1. Overview

The prototype (Figure 1) can be divided into the following main components: (1) real-time speech-to-text engine, (2) bot NLU for intent detection and reply generation, (3) expressive TTS engine, and (4) bot middleware that connects all components and manages audio channel to the call center software.

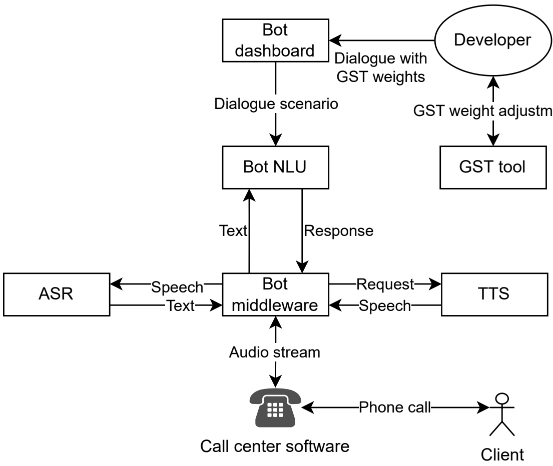

Figure 1: *Overview of the Spoken Dialogue System for Call Centers with Expressive Speech Synthesis.*

Additionally, the prototype includes (1) a bot dashboard or management system and (2) a global style token (GST) preparation tool. In the management system, a dialogue between a client and a bot is designed by defining a scenario, represented as a graph similar to a final state machine. When creating a new bot, the bot designer in advance defines the possible topics, questions, and replies. The bot designer can also choose the appropriate sentiment for each phrase by providing GST weights.

The GST tool provides a simple interface that allows the bot designer to tailor each response's sentiment by manually adjusting the GST weights and their respective prosodic features or by uploading a reference audio file containing the required emotion. When the appropriate GST weights are determined, they are exported to the bot dashboard and assigned to the corresponding reply nodes in the scenario.

2.2. Speech Recognition

The Latvian real-time speech recognition is based on the online speech decoder from the Kaldi toolkit [2]. The architecture of the acoustic model is the TDNN deep neural network trained with the LF-MMI loss function [3]. For language modeling, 3-gram LM is currently used considering recognition quality and latency trade-off. Subword BPE [4, 5] units are used instead of full words to tackle the morphology and inflective nature of the Latvian language.

2.3. Intent Detection

Our intent detection system is based on fastText [6] word embeddings and a convolutional neural network classifier. Such a model has low inference latency and is capable of achieving state-of-the-art results [7]. We have trained separate language-specific models for languages commonly spoken in Baltic countries: Estonian, Latvian, Lithuanian, English, and Russian. When creating a dialogue scenario, the developer provides examples of user inputs and intents. Each intent must have at least 5 user input examples. These examples are then used as training samples to fine-tune the intent detection model.

2.4. Speech Synthesis

For the bot response synthesis, we use the expressive TTS prototype from our previous work [1]. It is based on Tacotron 2 [8], one of the current state-of-the-art TTS models that produce mel-spectrograms based on grapheme or phoneme input. While it can produce natural-sounding speech, it has limited ability to model and control prosodic information. Therefore, we use an extension to the Tacotron's architecture proposed by [9], which aims to capture the non-textual information of the speech in a set of trainable embeddings called Global Style Tokens (GSTs).

The GSTs are trained in an unsupervised manner, based on the Tacotron 2 decoder's reconstruction loss. The stylistic information of a speech sample is captured by a reference encoder, which forms a reference embedding from the sample's mel-spectrogram. The reference embedding is compared to each style token with an attention mechanism, which measures the similarity between the reference embedding and the style token. The similarity measures are used to form a weighted sum of the style tokens, resulting in a style embedding that is concatenated to the text encoder output. Thus, the mel-spectrogram synthesis can be conditioned on both textual and stylistic information.

2.5. Bot Middleware

The bot middleware is based on a Tornado WebSocket server. When a client calls the bot, the call center software initiates a WebSocket connection with the bot middleware. The connection serves as an audio stream for receiving input and returning output to the phone line.

First, the incoming audio is streamed to the ASR. ASR processes the audio in real-time, and as soon as the final hypothesis is formed, it is immediately passed to the bot platform for intent detection.

After retrieving an appropriate response from the bot platform, the response text is passed to the expressive TTS service. The audio is synthesized with a 24kHz sampling rate and needs to be downsampled before it can be sent to the phone line.

One challenge for the spoken dialogue system is to ensure real-time communication between the client and the bot. All components need to ensure low latency. All processing steps are done asynchronously to avoid blocking the input/output stream between the bot middleware and the call center. Processing and playback of all previous requests is stopped once a new request from the client is received to maintain bot responsiveness. The output audio is streamed in small chunks with a slight delay in-between the chunks to avoid audio build-up on the call center side and ensure that the playback can be stopped without significant delay.

3. Conclusions

In this demonstration paper, we presented a prototype of a fully voice-enabled bot solution. This allows the bot to be integrated into call centers or to be used in other scenarios, where voice communication is preferable. While our current ASR, NLU, and TTS models can achieve reasonable results, there is still room for research and improvement. The latency of the whole solution also can be further improved. The prototype is primarily designed for the Latvian language; however, for the purpose of demonstration, we plan to add English ASR and TTS to the prototype.

4. Acknowledgements

The research has been supported by the European Regional Development Fund within the research project "AI Assistant for Multilingual Meeting Management" No. 1.1.1.1/19/A/082.

5. References

[1] D. Nicmanis and A. Salimbajevs, "Expressive Latvian Speech Synthesis for Dialog Systems," in *Proc. Interspeech 2021*, 2021, pp. 3321–3322.

[2] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[3] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-Sept, 2016, pp. 2751–2755.

[4] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1715–1725.

[5] P. Smit, S. Virpioja, M. Kurimo *et al.*, "Improved subword modeling for wfst-based speech recognition." in *INTERSPEECH*, 2017, pp. 2551–2555.

[6] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the association for computational linguistics*, vol. 5, pp. 135–146, 2017.

[7] K. Balodis and D. Deksne, "Fasttext-based intent detection for inflected languages," *Information*, vol. 10, no. 5, 2019. [Online]. Available: https://www.mdpi.com/2078-2489/10/5/161

[8] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[9] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.