INTERSPEECH 2021: Show & Tell Contribution
30 August – 3 September, 2021, Brno, Czechia

Expressive Latvian Speech Synthesis for Dialog Systems

Dāvis 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

To fully enable spoken human-computer interaction, the text-to-speech (TTS) component of such a system must produce natural human-like speech and adjust the prosody according to the dialog context.

While the current publicly available TTS services can produce natural-sounding speech, they usually lack emotional expressiveness.

In this paper, we present an expressive speech synthesis prototype for the Latvian language. The prototype is integrated into our chatbot management system and enables bot designers to specify the stylistic information for each bot response, thus making the interaction with the chatbot more natural.
Index Terms: expressive speech synthesis, human-computer interaction, dialog system, chatbot

1. Introduction

As the pandemic broke out, it reshaped the world and shifted many aspects of our lives to online. For example, in Latvia, this resulted in a surge of interest in chatbots - virtual assistants that typically help users to navigate websites of government agencies and private companies. Most of such chatbots are text-based dialog systems, however both in Latvia and worldwide [1] there is a growing interest in voice-based applications, thanks to their contactless interaction method, which is crucial during the COVID-19 pandemic.

While chatbots can be easily integrated with the current publicly available TTS services to produce natural-sounding speech, in many cases, such solutions lack expressiveness. Expressivity, i.e., intonation, rhythm, and stress, are an essential part of human verbal communication. When we talk with each other, we change our expression to show empathy or convey additional context to the information. Similarly, different expressions can help a chatbot, e.g., emphasize important safety information or appropriately respond to a frustrated customer support inquiries.

In this paper, we present an expressive Latvian speech synthesis prototype for dialog systems. The prototype is integrated with our proprietary chatbot management system and enables bot designers to adjust the sentiment of each bot response. The prototype source code is avalaible at https://github.com/tilde-nlp/pip2-expressive-speech-synthesis-for-dialogs.

2. Method

2.1. Semi-automated speech annotation

Given the lack of annotated expressive speech data in the Latvian language, we implemented a semi-automated speech annotation tool for preparing TTS training data with the following pipeline (inspired by [2]):

- Source text normalization. The text should match the audio as close as possible. Automatic number rewriting is applied, fragments, which are not read by the speaker, are filtered using regular expressions;
- Fuzzy alignment. The audio is aligned with the reference text using a fuzzy automatic speech recognition (ASR) alignment method, as the ASR force alignment is unable to process long segments that have large differences between the audio and reference text. In result, a list of recognized words, timestamps, and confidence scores are obtained. The words with high confidence are used as ”anchor” points for the segmentation;
- Segmentation. The text and the audio is divided into smaller utterances based on punctuation and pause duration between anchor points;
- Force alignment. ASR force alignment is performed for each segment. Unalignable segments are filtered out.

2.2. Data

Our baseline dataset was specifically recorded for the training of an end-to-end TTS model for the telecommunication domain. It comprises hand-picked phrases, advertisement fragments, and blog articles, with a total size of 8604 utterances, 12.56 hours. However, it lacks the expressiveness and emotional variability as the speaker was explicitly asked to deliver utterances consistently and in a neutral expression.

In order to get training samples for expressive speech, we selected three audiobooks and processed them using the pipeline described in the previous subsection:

- An autobiographical book, containing mostly formal language related to the business domain. In total: 3319 utterances, 5.73 hours;
- A novel, containing colloquial language. In total: 3830 utterances, 4.79 hours;
- A novel with an increased frequency of fictional and translated words, containing phonemes, which are uncommon in the Latvian language. In total: 3009 utterances, 4.65 hours.

Moreover, we asked the speaker of the first audiobook to record an additional set of hand-picked phrases in neutral, happy, sad, and insistent expressions. In total: 5957 utterances, 13.31 hours.

All described datasets were combined into a single multi-speaker dataset to cover different domain texts, increase the

overall prosody variety and the total training data amount. All datasets are recorded by different speakers, all of whom are female. All audio files were converted to 16-bit 24kHZ mono PCM for the model training.

2.3. Global Style Tokens

Our speech synthesis model is based on Tacotron 2 [3], one of the current state-of-the-art TTS models that produces 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 extension to the Tacotron's architecture proposed by [4], 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.

3. Expressive Speech Synthesis for Dialogs

Expressive Latvian Speech Synthesis for Dialogs (Figure 1) can be divided into 4 separate components: (1) Bot dashboard or management system, (2) GST preparation tool, (3) Bot and (4) Expressive TTS engine.

In our chatbot management system, a dialog between a user and a bot is based on a predefined 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. Hence, the bot designer can manually choose the appropriate sentiment for each phrase.

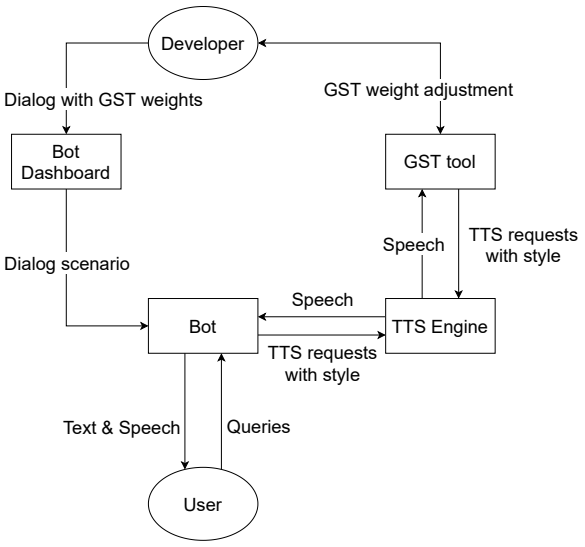

Figure 1: *Overview of the Expressive Speech Synthesis system for Dialogs.*

When the bot scenario is defined, the next step is to find the appropriate style token combination for each sentiment. Since the GSTs are learned in an unsupervised manner, it is often difficult to interpret the role of each token. For example, one token can be responsible for the pause duration between words, while a different token can modify the overall pitch of the speech.

The GST calculation can be conditioned in two ways - by passing a dictionary of manually selected token weights or with an audio file. In the dictionary input case, we select the specified token values, multiply them with the respective weights, and sum them together, forming the final style embedding. In the audio input case, the token weights are calculated in the same way as during the training - the audio file is encoded by the reference encoder and compared to each style token with an attention mechanism, again forming the weights that are used for the weighted sum.

The GST tool incorporates both of these methods in 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 chatbot management system and assigned to the corresponding reply nodes in the scenario. During the synthesis of a reply, the bot makes a request to the TTS service and passes both the text and the GST weights. The TTS service is based on a master server, which handles the requests and responses, and delegates the synthesis to a pool of worker processes.

4. Conclusions

In this demonstration paper, we presented a prototype of Expressive Speech Synthesis for Dialog Systems. While our current TTS model is able to achieve reasonable expressiveness, there is still room for research and improvement in terms of prosody control. The prototype is prepared for the Latvian language; however, the presented method can be applied to any other language.

5. Acknowledgements

This research has been supported by the European Regional Development Fund within the joint project of SIA TILDE and University of Latvia "Multilingual Artificial Intelligence Based Human Computer Interaction" No. 1.1.1.1/18/A/148.

6. References

[1] M. Webster, "Voice technology's role in our rapidly changing world: Adobe xd ideas," Oct 2020. [Online]. Available: https://xd.adobe.com/ideas/principles/emerging-technology/voice-technologys-role-in-rapidly-changing-world/

[2] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," 04 2015, pp. 5206–5210.

[3] 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.

[4] 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.