Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6633–6637
Marseille, 11–16 May 2020

Development and Evaluation of Speech Synthesis Corpora for Latvian

Roberts Darģis[1], Pēteris Paikens[1], Normunds Grūzītis[1], Ilze Auziņa[1], Agate Akmane[2]
[1]Institute of Mathematics and Computer Science, University of Latvia, 29 Raina Blvd., Riga, Latvia
[2]Riga East Clinical University Hospital, 2 Hipokrata Street, Riga, Latvia
roberts.dargis@lumii.lv

Abstract

Text to speech (TTS) systems are necessary for any language to ensure accessibility and availability of digital language services. Recent advances in neural speech synthesis have enabled the development of such systems with a data-driven approach that does not require significant development of language-specific tools. However, smaller languages often lack speech corpora that would be sufficient for training current neural TTS models, which require at least 30 hours of good quality audio recordings from a single speaker in a noiseless environment with matching transcriptions. Making such a corpus manually can be cost prohibitive. This paper presents an unsupervised approach to obtain a suitable corpus from unannotated recordings using automated speech recognition for transcription, as well as automated speaker segmentation and identification. The proposed method and software tools are applied and evaluated on a case study for developing a corpus suitable for Latvian speech synthesis based on Latvian public radio archive data.

Keywords: speech corpus development, text-to-speech, speech recognition, speaker segmentation

1. Introduction

The popularity of speech synthesis as a topic in natural language processing has significantly increased after the publication of results by DeepMind (van den Oord et al., 2016), Baidu (Arik et al., 2017) and Google (Wang et al,, 2017; Shen et al., 2018), demonstrating the ability to create natural sounding speech with neural methods. While the new methods themselves are mostly language agnostic, extending these text to speech (TTS) systems for a particular language requires a specialized speech corpus for that language.

Training these currently mainstream speech synthesis methods from scratch to adequate quality require about 30 hours of good quality audio recordings from a single speaker in noiseless environment, and an accurate transcription of these recordings. For less resourced languages, obtaining such a corpus is the major obstacle to development of TTS solutions.

In this paper an unsupervised approach for such corpus creation is presented, using ASR and speaker segmentation as main components.

2. Related Work

Before this research, a suitable corpus was not available for Latvian: there was a sizeable transcribed speech corpus (Pinnis et al., 2014) developed for automated speech recognition (ASR) purposes, but, because of ASR requirements, the corpus intentionally had a wide diversity of speakers, and did not have sufficient data from any single speaker to train a good TTS system.

The usual approach to develop such a corpus involves narrators who record especially selected text segments in studio conditions. This approach has been used in multiple popular speech corpora of other languages, such as CMU ARTIC (Kominek et al., 2017) and CSTR VCTK Corpus (Veaux et al., 2017).

The other alternative to obtain high quality data would be to transcribe existing audio recordings, but that requires a lot of manual work. Approximately 10 hours of work are required to transcribe one hour of raw audio from scratch, with some speed up possible by editing automatic transcription if ASR is available for the target language.

For the purposes of this Latvian TTS case study, both these approaches were unsuitable due to resource constraints.

Another alternative that has been used for similar goals (Székely et al., 2012) would be to align existing audio data where known corresponding text is available. A typical data source for this approach is audiobooks. We performed some preliminary experiments with audiobook data, however, the application of these methods was limited by copyright issues, the difficulty to obtain a sufficiently large corpus from a single narrator, and the genre specific properties of rhythm and intonation, as a "storytelling voice" was subjectively considered less suitable for many TTS applications.

A more complex approach to voice database creation has been used in creation of VoxCeleb database (Chung et al., 2018). They created an audio-visual dataset consisting of short clips of human speech, extracted from interview videos uploaded to YouTube. Due to the lack of audio transcripts, this corpus is mainly used for development of speaker verification and speaker recognition systems.

This paper builds upon these ideas and leverages publicly available solutions for ASR, diarization and speaker identification in order to obtain a speech corpus suitable for TTS needs in an unsupervised manner from raw, unannotated audio files.

3. Source Dataset

The first step in a corpus development is the selection of source data. Speech synthesis generally requires high quality studio audio recordings. The audio recordings should be:

- Noiseless – without any background noise, other conversations or music.
- From a native speaker with a good, clear pronunciation that matches the prescribed language standards without any slips of the tongue or mispronunciations.
- Fluent – it should be prepared, non-spontaneous speech without any unnecessary breaks, non-lexical vocables, fake starts (such as words and sentences that are cut off mid-utterance, phrases that are restarted or repeated and repeated syllables, fillers (such as "huh", "uh", "erm", "um", "well").

For this case study, we chose the broadcast news from Latvian national public radio broadcasting network

"Latvijas radio" as the source for audio recordings. We considered this as the most suitable source for the following reasons:

- "Latvijas Radio" archive has a large quantity of data accessible to the general public. It has an online archive dating back from 2003, including a news broadcast in the beginning of every hour for approximately 5 minutes.
- Their news recordings have no background music. This is an important factor, as all the other major radio broadcasting networks in Latvia have some background music when news is being read.
- The news are read from a prepared release, so the speech is fluent; unlike in speech from talk shows, interviews and other radio programs which contain a lot of spontaneous speech with non-lexical vocables multiple people talking over each other.
- The narrators are from a limited set of professional news anchors, so we can obtain large quantities of source data from the same speakers and the public radio traditionally places great importance in proper pronunciation in their selection.

From this source we managed to obtain approximately 500 hours of unfiltered Latvian broadcast news recordings from various speakers.

4. Corpus Processing Pipeline

Development of an usable corpus from the raw data obtained was performed with a processing pipeline consisting of the following steps:

1) Automated speech recognition – all audio files were processed using an existing ASR solution.
2) Speech segment extraction – based on the ASR results, the news broadcast was segmented, extracting parts that contain pure speech data that the ASR could recognize with high confidence.
3) Speaker segmentation – diarization of the valid speech segments in order to split them into segments of a single speaker.
4) Speaker clustering – attempt to cluster the single speaker segments, to obtain 'voice signatures' of the most popular speakers in the source data.
5) Speaker identification – the voice models obtained in previous step were used to select the final corpus of a particular speaker.

4.1 Speech Recognition

Automated speech recognition is the only language dependent step in the proposed corpus creation pipeline. In addition to accuracy, a very important ASR aspect for this corpus creation is the need for consistent word confidence scores, as this metric will be used to discriminate between 'clean' and noisy speech segments. As long as the word confidence scores are reliable, we can tolerate low recognition accuracy or ASR that is very sensitive to noisy input data or speaker qualities. The system will select the subset of data that the ASR can recognize, so these factors will just reduce the proportion of the data that can be kept for further steps in corpus creation, making this approach usable even if the ASR systems for target language are not particularly accurate compared to English and other major languages.

However, the need to obtain detailed confidence score information means that you need full access to the ASR system internals, as some commercial 'black-box' services only provide the output text. For the Latvian case study, we used an open-source ASR system developed in the Horizon 2020 project SUMMA[1] for Latvian speech recognition. The particular speech recognition model obtained a word error rate (WER) for clean audio recording of 12%, and was based on Kaldi speech recognition toolkit (Povey et al., 2011).

4.2 Speech Segment Extraction

The next step is filtering speech segments based on the ASR confidence score. An important parameter choice for this process is determining the threshold for word confidence score above which the speech segments will be selected as suitable, based on correctly recognized words (words above threshold) and pauses between them. For a segment to be suitable for the speech synthesis corpus, it should contain only words with confidence above the threshold, and we also impose an empirical limit on segmentation that the pause before and after the segment should be at least 0.2 seconds.

To determine the best threshold value, an automatic speech transcription was obtained for small speech corpus – about 30 minutes of audio for which we had human-verified correct transcriptions. Based on words that were above the confidence threshold, three parameters were computed as shown in Figure 1:

- Percentage recognized correctly – percentage of words that were recognized correctly.
- Duration of words (in percentage) – total duration of the words that have confidence above threshold, divided by duration of all files.
- Duration of words in segments (in percentage) – total duration of words in the speech segments divided by duration of all files.

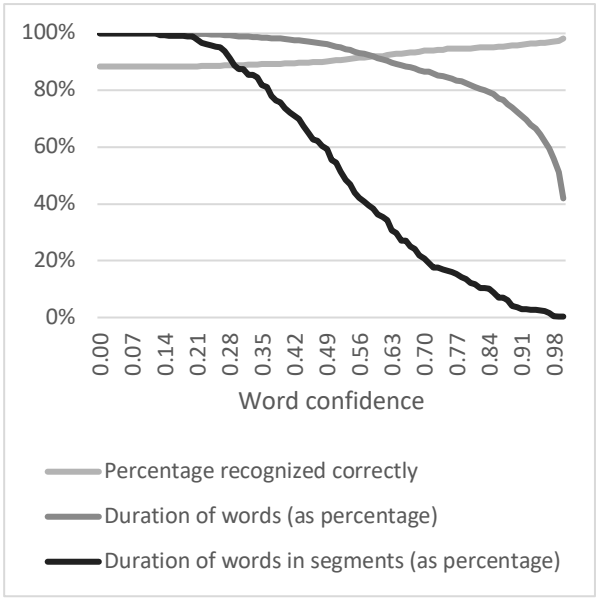

Figure 1: Speech segment extraction characteristics.

[1] SUMMA project: http://summa-project.eu/

The perfect threshold value would be the lowest confidence score where word recognition precision reaches 100%. Unfortunately, the word confidence score is not so consistent – even at the maximum word confidence score 1.00, the percentage of words recognized correctly is 98.21%. In addition, only 41.93% of words have this maximum word confidence score, and if we would pick that as the threshold, only 0.33% of the total source data would be kept, as almost all speech segments (separated on each side with at least 0.2 seconds of silence) include at least one word with a lower confidence score. Such low percentage is not practical, so a compromise is needed.

One difference between ASR and TTS that matters for this task is the homophone words (spelling variations with identical pronunciation) that are counted as errors in ASR, but they are completely appropriate for the speech synthesis. There are even some common phrases that are pronounced the same way because of sound reduction, for an English example: 'let her' and 'led her'; 'but her' and 'butter'. For a speech synthesis corpus, we can safely select such variations, as those errors do not affect TTS accuracy. In some sense, we care whether the phonetic or audio part of the ASR system has made a mistake or was not able to recognize a word, but the performance of ASR language or spelling model is not relevant.

The final word confidence threshold was chosen empirically, based on the abovementioned data and repeated experiments, manually going through the aligned transcription at various word confidence levels and judging whether the recognized words seemed acceptable for TTS needs. The ASR word confidence threshold chosen in this manner was 0.70. At this level 93.88% of words were recognized correctly and the percentage of audio kept increased to 20.71% of all the source data. While the majority of data was discarded, it is important to note that a large quantity of data even with a perfect ASR would have to be discarded as it contained silence, music and introduction sounds, or some voice segments over a noisy background.

4.3 Speaker Segmentation

As the initial segmentation is done based on pause length, it is possible that one extracted speech segment can contain speech from multiple speakers. In order to create speech segments that contain speech from only individual speakers, we use the speaker diarization system from LIUM_SpkDiarization tools (Rouvier et al., 2013). This speech segmentation gives the most probable time intervals for individual speakers. If we want to acquire good segments for the corpus, the speech can only be split in silences, so this information about individual speaker segments is merged with the time positions of word boundaries from the ASR system to separate the actual speaker segments that could be used in the corpus.

4.4 Speaker Clustering

Speaker clustering groups speech segments together by speaker across multiple files. The clustering methods that we used are too slow to be used directly on the full set of speech segments, so a subset of segments is selected to build speaker models ('voice signatures') that will be used for speaker identification in the next step of the corpus development pipeline. The subset selection is based on two assumptions – that each 5 minute news program is most likely spoken by a single main news anchor, and the longest individual speaker segment most likely belongs to the main news anchor. Although these assumptions do not always hold, it is enough to ensure that we get a representative sample of voice segments that will include multiple examples from all the most frequently heard news anchors. For the actual clustering, we used 600 random speech segments from the previously selected subset. The automated clustering system considered that there might be 246 speakers in these 600 speaker segments. Analyzing the frequency data clustering result (Table 1) illustrates that there are 3 main news anchors (S001 – S003) over the time span of this archive, which was verified with manual review of the sample data. Although it is unlikely that there actually were so many different speakers in the subset, as a random sample of the less frequent speaker IDs often revealed noisy or non-speech sound segments that prevent proper identification, the duration of the clustered segments for these top 3 speakers was sufficient to train a speaker model for speaker identification.

Speaker ID	**Number of segments**	**Percentage of segments**	**Time** (in seconds)
S001	190	31.25%	1476
S002	87	14.31%	517
S003	58	9.54%	491
S004	5	0.82%	53
S005	3	0.49%	37
S006	3	0.49%	34
…	…	…	…
S245	1	0.16%	1
S246	1	0.16%	1

Table 1: Speaker clustering result.

4.5 Speaker Identification

With speaker models developed in previous step, speaker identification was done on all individual speaker segments using the speaker identification system from the same diarization tools as in the previous step (Rouvier et al., 2013). This allows us to obtain a similarity score comparing each segment with each speaker in the previously built speaker model.

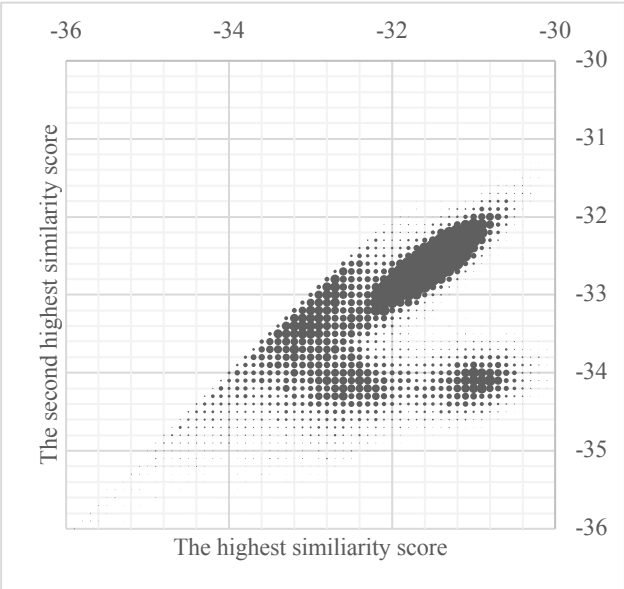

Figure 2: Speaker similarity score relation.

Because the speaker identification calculates scores only for the three individual voices included in the speaker model, we need to also determine if this segment is actually not spoken by one of these three news anchors.

To figure out a way how to detect if the speaker is not in the speaker model, all of the speech segments were scored against the speaker model, the scores were sorted by value (highest first) and the first two values were plotted in the bubble graph, as shown in Figure 2.

Based on the graph and some experimentation, the identified speaker will be considered as actually being from the speaker model only if the highest speaker score is above -32.5 and the difference between the second highest score will be at least 1.0. According to these criteria, a total of 44% (60 out of 135 hours) speech segments were identified as belonging to the one of three speakers – 23% (32 hours) to the first speaker, 14% (19 hours) to the second speaker and 7% (9 hours) to the third most frequent speaker.

5. Evaluation of the Corpus

In order to evaluate the newly created speech corpus we trained two text-to-speech systems – a parametric TTS system, and a solution based on Tacotron 2 (Shen et al., 2018) with WaveGlow (Prenger et al., 2018) system.

The parametric speech synthesis model was trained using Merlin toolkit (Wu et al., 2016) – a set of software for building Deep Neural Network models for statistical parametric speech synthesis.

For the Tacotron 2 approach, the implementation by Nvidia[2] was used to generate the mel-scale spectrograms of the generated speech. After that, a flow-based generative network for speech synthesis called WaveGlow[3] was used to generate the actual audio waveform from these mel-scale spectrograms. The learned attention alignment shown in Figure 3 demonstrates that the created corpus is viable for the development of speech synthesis systems.

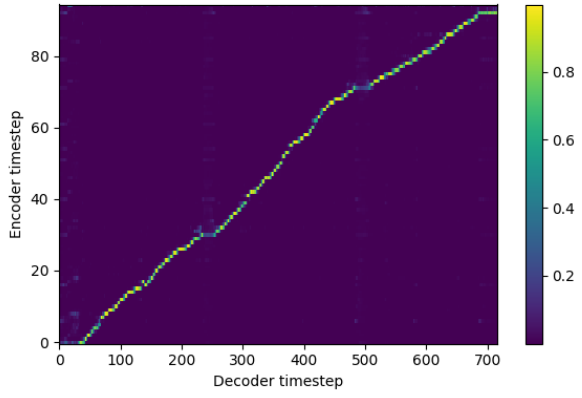

Figure 3: Learned attention alignment.

Using these two text-to-speech models, a sample audio set was generated for 100 speech segments (i.e., actually spoken by the same new anchor) that were withheld from the corpus. The generated sample set alongside the original reference audio was made available to the evaluators. For comparison purposes, we also included samples generated with an older concatenative Latvian TTS (Pinnis and Auziņa, 2010), which is the only currently published TTS solution for Latvian.

Each sample was reviewed by 10 evaluators who independently rated the quality with an opinion score from 1 to 5 (from bad to excellent). The overall mean opinion score (MOC) for each system is shown in Table 2. The best MOC across all segments and evaluators is 3.7, with a substantial improvement over the earlier concatenative TTS system. This is evidence that the corpus is viable for training text-to-speech systems, and thus validates the unsupervised method for corpus collection.

System	MOC
Concatenative	1.9
Parametric	3.2
Tacotron 2	3.7

Table 2: Mean opinion scores.

6. Conclusions and Further Work

In this paper we presented an unsupervised approach to obtain a speech corpus suitable for developing text to speech systems. Compared to other methods, this method does not require a properly transcribed text to be available beforehand and requires almost no manual work, and can be useful for rapid corpus development if an ASR system is available for the target language.

To demonstrate the viability of this approach, a speech corpus for Latvian was created and evaluated. The corpus is not yet publicly released due to unclear legal copyright aspect of the source audio recordings, but is available for research purposes. The corpus is based on approximately 500 hours of broadcast radio news recordings with multiple speakers, out of which a 60 hour corpus of selected high quality speech segments was obtained, containing speech from 3 different speakers. The majority of discarded data (79%) was due to the shortcomings of automatic speech recognition quality. By improving the ASR quality, more usable data could be extracted from the same source material. Another 9% of the initial data (56% of the data remaining after ASR filter) was discarded due to the speaker identification process.

Based on this corpus, two different TTS models were created, and user evaluation was conducted with ten separate evaluators. The mean opinion score was 3.7 out of 5. One of these models was further developed in a usable TTS system (Darģis and Auziņa, 2018), demonstrating that the corpus developed using this unsupervised approach is viable for development of text-to-speech systems.

Despite only 12% of the original data was used in the corpus, the amount was sufficient for practical purposes. As showed by multiple text-to-speech developers, 30 hours of data is enough to create a state-of-the-art text-to-speech system, and further evaluation would be needed on how and if extra training data would improve the quality of resulting TTS models.

7. Acknowledgements

This work has received financial support from the European Regional Development Fund under the grant agreements No. 1.1.1.1/18/A/153.

[2] NVIDIA's Tacotron 2 implementation: https://github.com/NVIDIA/tacotron2

[3] NVIDIA's WaveGlow implementation: https://github.com/NVIDIA/waveglow

8. Bibliographical References

Arik, S.O., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., Sengupta, S. and Shoyebi M. (2017). Deep voice: Real-time neural text-to-speech. arXiv preprint arXiv:1702.07825 (2017).

Chung, J.S., Nagrani, A., and Zisserman, A. (2018). VoxCeleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622

Dargis, R. and Auzina, I. (2018). Towards a Modern Text-to-Speech System for Latvian. In Human Language Technologies – The Baltic Perspective, Frontiers in Artificial Intelligence and Applications vol. 307, pages 26–29, IOS Press.

Kominek, J., Black, A. W. and Ver, V. (2003). CMU ARCTIC databases for speech synthesis.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. arXiv preprint, arXiv:1609.03499.

Pinnis, M., Auziņa, I., Goba, K. (2014). Designing the Latvian speech recognition corpus. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC), pages 1547–1553, European Language Resource Association (ELRA).

Pinnis, M. and Auziņa, I. (2010). Latvian text-to-speech synthesizer. In Proceedings of the 2010 conference on Human Language Technologies–The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT 2010, pages 69–72. IOS Press.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hanneman, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi speech recognition toolkit. In IEEE 2011 workshop on automatic speech recognition and understanding. IEEE Signal Processing Society.

Prenger, R., Valle, R. and Catanzaro, B. (2018). Waveglow: A flow-based generative network for speech synthesis. arXiv preprint arXiv:1811.00002.

Rouvier, M., Dupuy, G., Gay, P., Khoury, E., Merlin, T., and Meignier, S. (2013). An open-source state-of-the-art toolbox for broadcast news diarization. In proceedings of Interspeech 2013.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, RJ, Saurous, R. A., Agiomyrgiannakis Y., and Wu, Y. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4779-4783). IEEE.

Székely, É., Cabral, J P., Abou-Zleikha, M., Cahill, P., Carson-Berndsen, J. (2012) Evaluating expressive speech synthesis from audiobooks in conversational phrases. Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pages 3335–3339, Istanbul, Turkey, May. European Language Resource Association (ELRA).

Wang, Y., Skerry-Ryan, RJ, Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R. and Saurous, R.A., (2017). Tacotron: Towards End-to-End Speech Synthesis. arXiv preprint arXiv:1703.10135.

Wu, Z., Watts, O., & King, S. (2016). Merlin: An open source neural network speech synthesis system. *Proc. SSW, Sunnyvale, USA*.

9. Language Resource References

Veaux, C., Yamagishi, J. and MacDonald, K. (2017). CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit. University of Edinburgh. The Centre for Speech Technology Research (CSTR). https://doi.org/10.7488/ds/1994.