



A Method for Selection of Phonetically Balanced Sentences in Read Speech Corpus Design

Jerneja Žganec Gros
Faculty of Industrial Engineering
Novo Mesto, Slovenia
 0000-0001-5011-8486

Boštjan Vesnicer
Alpineon Ltd
Ljubljana, Slovenia
bostjan.vesnicer@alpineon.si

Simon Dobrišek
Faculty of Electrical Engineering
University of Ljubljana
Ljubljana, Slovenia
 0000-0002-9130-0345

Abstract—Sentence selection for speech prompts plays an important role in the process of designing a speech corpus of read speech, both for speech recognition and speech synthesis. The presented method for selecting a phonetically balanced subset of sentences from a larger sentence set can also be used in linguistic research. Especially in research settings, where the n-gram distribution of letter or phone constituents in the target sentence set is expected to reflect a preselected reference or target distribution.

Keywords—speech technologies, speech corpora, text corpora, sentence selection

I. INTRODUCTION

In the development of speech technologies, speech corpora play a key role in training and testing acoustic models, both in speech recognizers and speech synthesizers. Especially when it comes to automatic speech recognition, we want these corpora to be as large as possible.

Recent advances in speech technologies, especially speech recognition, have accelerated the creation of extensive speech corpora, such as the Mozilla Common Voice dataset¹ for training speech models [1], the VoxPopuli Corpus² [2], and the multilingual LibriSpeech collection³ [3]. In order to cover open vocabulary tasks, various types of linguistic units, apart from standard word units have been examined, such as morphemes, syllables, phonemes [4] and data-derived linguistic units, like byte-pair encoding [5], [6], and Morfessor variants [7].

When designing and building a speech corpus, we need to keep in mind the purpose and type of use of the final speech recognizer. This implies what kind of speech data we wish to include into the speech corpus in order to efficiently train the acoustic model.

In the paper we describe the process of sentence selection for speech prompts in the read speech part of the speech corpus in the project Development of Slovene in a digital environment (DSDE)⁴, designed to train a general-purpose speech recognizer.

II. SPEECH CORPUS DESIGN

The creation of speech corpora is a time-consuming and expensive task, so the size of these corpora should be limited to what is still acceptable. Fortunately, the relationship between training speech corpus size and the speech recognition accuracy is not linear, but is approaching logarithmic [8], which means that from a certain corpus size, increasing the size of the training speech corpus acquired under the same acoustic conditions no longer pays off. For

example, Wu et al. showed that training a HMM speech recognizer can achieve comparable results with a properly selected subset of sentences from a larger speech corpus than training the speech recognizer on the entire sentence set in the speech corpus [9].

Given the limited resources for the speech corpus design, a careful selection of sentence prompts to be read by the selected speakers is all the more important.

III. SENTENCE SELECTION FOR SPEECH PROMPTS

The sentences that the speakers will be prompted to read and record can be selected at random from a larger set of sentence candidates, as in the Mozilla Common Voice project [1].

However, this is not optimal [9], [10], as shown in a study [11], where Gouvêa et al. compared the impact of different sentence selection criteria for speech recognition training data on the accuracy of the resulting speech recognizer. Research has shown that the best results are obtained when the distribution of phoneme n-grams in the selected training data matches well enough with the distribution of n-grams in the test data, thereby mimicking general utterances in a given language. Similar findings were made in a related study by Kleynhans and Barnard [12]. They used triphones for basic phonetic units. As the main conclusion, they pointed out that matching the "natural" or actual distribution is the most suitable criterion for choosing a representative sentence set.

The question arises as to how to choose such a subset of sentences from a larger set of sentences, which will represent the optimal choice according to a selected criterion.

It turns out that the problem of optimal sentence selection is very similar to the coverage problem [13] or its variation, i.e., the maximum coverage problem. Since we know that this problem falls into the class of problems for which an effective algorithm is not available, we need to resort to approximate algorithms in the form of greedy methods. The basic idea is simple. We start with an empty set of selected sentences, then in each step we select the most optimal sentence from the whole sentence corpus according to the current coverage conditions and we add this sentence to the set of selected sentences.

A. Sentence Selection Criteria

A prerequisite for optimal sentence selection is to determine the criterion by which sentences are chosen. In general, we want the speech corpus to be comprised of such a collection of read sentences so that the representation of phonetic units in the collection is either as uniform as possible (phonetically rich) or that it matches the actual frequency

¹ <https://commonvoice.mozilla.org/en/datasets>

² <https://github.com/facebookresearch/voxpathuli>

³ <http://www.openslr.org/94/>

⁴ <https://www.cjvt.si/rsdo/en/project/>

distribution of these units in natural speech, which means that it will be phonetically balanced, like the Harvard sentence corpus for English⁵ and its derivatives for other languages [13], [14], [15], and [16]. The first criterion is more appropriate when choosing sentences for speech synthesis, and the second for sentences for speech recognition.

A lot of research deals with similar topics [10], [17], [18], [19], and [20]. The sentence selection methods these authors propose differ in details only. Most methods start with an empty set of selected sentences. In further steps they choose a sentence from the entire sentence set, that is most relevant according to the selected criterion and they add this sentence to the set of selected sentences [21]. Others choose different tactics and instead of gradually adding sentences, they first place all sentences in the set of selected sentences and then gradually remove the sentences from that set until they reach the target set size of the set of selected sentences [22].

There is a third strategy which applies choosing sentence pairs. At each step the first sentence is selected from the set of selected sentences and the second from the set entire sentence set. The sentences are switched between the two sets in case the replacement is favourable according to the chosen criterion [23].

B. Sentence Selection Method

In the DSDE project our task was to select a target number of sentences out of a large number of sentences comprising several million sentences of the Slovene Gigafida 2.0 text corpus⁶ [24].

For the selection of sentences, we propose a novel method that represents a modification of the procedure by selecting sentence pairs. Because sentences are pooled from a large pool of sentences, not all sentences can be selected in a single step. Initially, sentences for the selected sentence set are chosen at random. In later steps of the sentence selection procedure, random pairs of sentences are selected (the first sentence from the set of selected sentences, the second sentence from the entire sentence set) and the sentence switching is performed only if it pays off according to the selected criteria. This means that the value of the criterion function increases after every sentence switch. And how do we determine in an iteration of the algorithm, that switching sentences between the two sentence sets pays off?

We are looking for such a target set of sentences in which the distribution of triphones will be as similar as possible to the actual distribution of triphones in the observed language. This means that we need to compare the two probability distributions. In mathematical statistics we can do this by calculating the relative entropy, also known as Kullback-Leibler divergence [11]. This divergence can be thought of as a measure of how the first probability distribution differs from the second – the reference one. When the divergence value is zero, both distributions are identical. When observing time sequences, this divergence can be simplified into the Jensen-Shannon divergence, which we chose in the sentence selection process as a measure of similarity. The Jensen-Shannon divergence is a smoothed and symmetrical derivative of the Kullback-Leibler divergence.

An important feature of the proposed sentence selection procedure is that the value of the criterion function in maximising the coverage increases monotonically towards a

local optimum, so the algorithm can be terminated at any time. To this end, we monitor the value of the criterion function, i.e., the Jensen-Shannon divergence, and when its value stabilizes, we stop the process. The required subset of relevant sentences has been selected.

IV. SENTENCE SELECTION METHOD FOR THE DSDE READ SPEECH CORPUS DESIGN

The sentence selection method presented was used to select a pool of sentence prompts for the read part of the speech corpus, which is being recorded in scope of the DSDE project. To train the speech recognizer, the DSDE project envisages recording half an hour of read text prompts per each of the 1000 speakers, which will consist of the selected sentences.

The entire collection of Gigafida 2.0 texts, with just under 60 million sentences, represented the reference set $S_{\text{Reference}}$ of sentences, on which we calculated the reference statistical distribution of phones. Due to Gigafida license restrictions, we were only able to perform the sentence selection from a limited pool of Gigafida sentences, which represented roughly 10% of the sentences added to Gigafida 2.0 in comparison to Gigafida 1.0. This part of the text corpus may be published under the CC-BY 4.0 license.

Speech recordings in the DSDE corpus of read speech will be equipped with a text transcript, which is represented by sentences from the text sentence prompts intended to be read by speakers. Due to budgetary restrictions, no pronunciation transcription has been foreseen for the read part of the DSDE speech corpus. Consequently, we further limited the selection of sentences to those where the pronunciation matches as much as possible the graphemic transcription of the sentence. Sentences containing words where the pronunciation differs from the spelling, e.g., abbreviations, acronyms, digits, foreign words, foreign proper names, etc. were excluded from the initial sentence set. We also attempted to automatically detect sentences that contain hostile, offensive, sexist or otherwise controversial content or spelling and grammatical errors and eliminated these sentences as well.

We call this set of sentences the initial sentence set S_{Initial} . The task was to select a phonetically balanced subset of sentences from the S_{Initial} pool of sentences so that they will yield about 500 hours of read speech by taking into account that speakers are reading the text sentence prompts at an average speaking rate.

Information on the actual frequency distribution of phonetic units in Slovene speech was estimated on the basis of a phonetic transcript of the entire Gigafida 2.0 text corpus. We call it the *Reference distribution*. The conversion from the graphemic annotation to the phonemic annotation was performed using automatic grapheme-phonemic conversion procedure for Slovene. From the phonemic transcripts of all sentences represented the reference set $S_{\text{Reference}}$ of sentences, we calculated the reference distribution of triphones, which was further used in the sentence selection procedure to measure the similarity between the target selected set of sentences S_{Selected} and the reference set of sentences $S_{\text{Reference}}$.

Since the sentence selection algorithm described in Chapter 3 can stop at a local optimum, it would be possible that the value of the criterion function depends much on the

⁵ <https://www.cs.columbia.edu/~hgs/audio/harvard.html>

⁶ <https://www.cjvt.si/en/research/cjvt-projects/gigafida-corpus/>

initial random selection of sentences in to the S_{Selected} set, which is performed during the process initialization in the first step of the sentence selection procedure.

To verify this, we repeated the sentence selection process several times with different randomly chosen initial sentence sets S_{Selected} . It turned out that the value of the criterion function converged to values that did not differ significantly from each other. This means that the various resulting sets of S_{Selected} are comparable according to the given criterion.

We confirmed this with experiments where we observed the values of the criterion function depending on the number of iterations of the sentence selection procedure. Figure 1 shows the values of the criterion function depending on the number of iterations of the sentence selection procedure. We see that in the first part the value falls rapidly and then stabilizes.

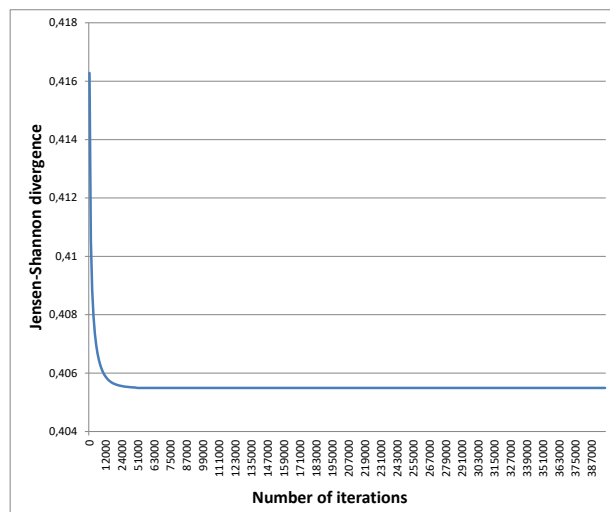


Fig. 1. The value of the criterion function, the Jensen-Shannon divergence depending on the number of pair-wise comparison iterations in the sentence selection procedure.

The value of the Jensen-Shannon divergence criterion function expresses how well the distribution of triphones in the selected sentence set S_{Selected} matches the distribution of triphones in the reference set $S_{\text{Reference}}$. The match between the two distributions can be shown by presenting the distribution graphically.

In Figure 2, the distribution of triphones in the reference set is presented by the blue curve. The distribution of triphones in the randomly selected set used to initialize the optimization process is presented by the orange curve. The curves were plotted by calculating the relative frequencies of individual triphones separately for each sentence set and by arranging these relative frequencies in descending order. The distribution of triphones in the selected set of sentences S_{Selected} , which was obtained with the proposed sentence selection method, is depicted by the green curve.

We can see that after stopping the optimization process, the distribution of triphones in the selected set fits very well the triphone distribution of the reference set, while the fit with a randomly selected set is much worse. The difference is more pronounced on the right part of the graph, which corresponds to less common triphones. This means that we prefer more frequent triphones by random selection, while the proposed

sentence selection methods select sentences in such a way that the triphone distribution match is good along the entire curve, even with rarer triphones.

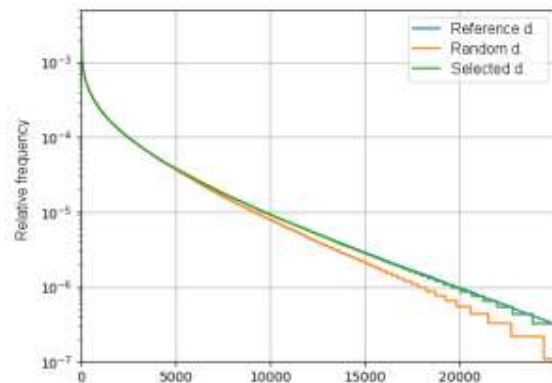


Fig. 2. Matching the distribution of triphones in the observed set of sentences with the reference distribution of triphones in the reference sentence set. The distribution of triphones in the reference sentence set is presented by the blue curve. The distribution of triphones in the randomly selected set used to initialize the optimization process is presented by the orange curve. The distribution of triphones in the selected set of sentences, which was chosen by the proposed sentence selection procedure, is presented by the green curve.

Finally, all sentences in the S_{Selected} underwent a final manual revision. In this process, sentences containing hostile, offensive, sexist or otherwise controversial or inappropriate content or language errors were excluded.

V. CONCLUSIONS

In the paper, we presented the process of selecting sentences of textual prompts for the read part of the Slovene speech corpus, which is being recorded in the DSDE project.

Speech database creation for training speech recognizers is a time-consuming and expensive task, therefore it is necessary to pay considerable attention to the design of the speech database. Our task was to select a set of most suitable sentences from an extensive collection of texts that would result in a target amount of speech recordings in the form read speech and will be most suitable for the intended specific purpose.

The question arises as to how to choose from a large number of sentences those sentences that will be most suitable for a specific purpose. What does suitability imply?

If we have two equally large subsets of sentences with which to train two speech recognizers, we can say that the more appropriate one is the set which yields a more accurate speech recognizer.

Such a suitability criterion is too impractical to be directly applicable, hence it was necessary to find a more practical alternative criterion. As an alternative criterion, we chose to match the distribution of triphones between the total set of sentences available and the selected set of sentences, thereby mimicking general utterances in a given language.

We assume that with two sentence subsets, the subset whose distribution of triphones better matches the distribution of triphones of the entire sentence set is more appropriate.

The presented method for selecting a phonetically balanced subset of sentences from a larger set of sentences can also be used in linguistic research. Especially in research settings, where the n-gram distribution of letter or phone constituents in the target sentence set is expected to reflect a preselected reference or target distribution.

The selected sentence set along with the corresponding audio recordings will be published under CC-BY 4.0 license in the CLARIN.SI language repository as soon as the speaker recording sessions and the validation of the recordings are concluded.

ACKNOWLEDGMENT

The research work has been performed in scope of the Development of Slovene in a Digital Environment project, funded by the Slovene Ministry of Culture and the European Regional Development Fund. The grapheme-to-phoneme transcription of the reference text has been performed by tools and language resources developed in scope of the OptiLEX project (L7-9406) and co-financed by the Slovene Research Agency.

REFERENCES

- [1] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Webe, "Common voice: A massively-multilingual speech corpus," In: Proceedings of LREC 2020: Twelfth International Conference on Language Resources and Evaluation: May 11-16, Marseille, France, 2020.
- [2] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, J., and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semisupervised learning and interpretation," In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 993–1003, Association for Computational Linguistics, August 2021.
- [3] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MIs: A large-scale multilingual dataset for speech research," In: Proceedings of Interspeech 2020, pp. 2757–2761. ISCA, 2020.
- [4] V. Papadourakis, M. Muller, J. Liu, A., Mouchtaris, and M. Omologo, "Phonetically Induced Subwords for End-to-End Speech Recognition," In: Proceedings of the Interspeech, pages 1992–1996, 2021.
- [5] Zeyer, A., Irie, K., Schluter, R., and Ney, H. (2018). "Improved training of end-to-end attention models for speech recognition," In: Proceedings of Interspeech, pages 7–11, 2018.
- [6] A. Laptev, A. Andrusenko, I. Podluzhny, A. Mitrofanov, I. Medennikov, and Y. N. Matveev, "Dynamic Acoustic Unit Augmentation with Bpe-Dropout for Low-Resource End-to-End Speech Recognition," Sensors (Basel, Switzerland), 2021.
- [7] S.-A. Gronroos, S. Virpioja, and M. Kurimo, "Morfessor EM+Prune: Improved subword segmentation with expectation maximization and pruning," In: Proceedings of LREC 2020: Twelfth International Conference on Language Resources and Evaluation: May 11-16, pages 3944–3953, Marseille, France, 2020.
- [8] R. K. Moore, "A comparison of the data requirements of automatic speech recognition systems and human listeners," In: Proceedings of the EUROSPEECH'93, 1993, 2582-2584.
- [9] Y. Wu, R. Zhang and A. Rudnicky, "Data selection for speech recognition," In: Proceedings of the ASRU, Pennsylvania, USA, 2007, 562-565.
- [10] J. P. H. van Santen and A. L. Buchsbaum: Methods for optimal text selection," In: Proceedings of the EUROSPEECH'97, Rhodes, Greece, 1997, 553-556.
- [11] E. Gouvêa and M. H. Davel, "Kullback-Leibler divergence-based ASR training data selection," In: Proceedings INTERSPEECH'11. Florence, Italy, 2011, 2297–2300.
- [12] T. Kleynhans and E. Barnard, "Efficient Data Selection for ASR," Language Resources and Evaluation 49/2 (2015), 327–353.
- [13] V. Aubanel, C. Bayard, A. Strauß, and J.-L. Schwartz, "The Fharvard Corpus: A Phonemically Balanced French Sentence Resource for Audiology and Intelligibility Research," Speech Communication, 124:68–74, 2020.
- [14] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, "Introduction to Algorithms," Cambridge, Ma.: MIT Press, 2020.
- [15] E. H. Rothauer, W. D. Chapman, N. Guttman, N., M. H. L. Hecker, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock, "Recommended Practice for Speech Quality Measurements," IEEE Transactions on Audio and Electroacoustics, 17(3):225–246, 1969.
- [16] A. Sfakianaki, G. Kafentzis, and Y. Stylianou, "The gr-Harvard corpus a greek sentence corpus for speech technology research and applications," In: Proceedings of the 3rd Summit on Gender Equality in Computing (GEC 2021), 2021.
- [17] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. Marino, and C. Nadeu, "Albayzin speech database: Design of the phonetic corpus" ,” In: Proceedings EUROSPEECH'03, Vol. 1, pp. 175–178, 1993.
- [18] V. Radova and P. Vopalka, "Methods of sentences selection for read-speech corpus design," In: Proceedings of the TSD '99, Springer Verlag, 1999.
- [19] Muljono, A. Harjoko, N. A. S. Winarsih and C. Supriyanto, "An evaluation of sentence selection methods on the different phone-sized units for constructing Indonesian speech corpus," International Journal of Speech Technology 23 (2020), 141–147.
- [20] H. Polat and S. Oyucu, "Building a Speech and Text Corpus of Turkish: Large Corpus Collection with Initial Speech Recognition Results," Symmetry 12 (2020), 290.
- [21] A. W. Black and K. A. Lenzo, "Optimal data selection for unit selection synthesis," In: Proceedings of the 4th Speech Synthesis Workshop SSW4, Perthshire, Scotland, 2001.
- [22] M. Rojc and Z. Kačič, "Design of an optimal Slovenian speech corpus for use in the concatenative speech synthesis system," In: Proceedings Language Resources and Evaluation (LREC), Athens, Greece, 2000, 1:321-325.
- [23] H. Kawai, S. Yamamoto, N. Higuchi and T. Shimizu, "A design method of speech corpus for text-to-speech synthesis taking account of prosody," In: Proceedings ICSLP, Beijing, China, 2000.
- [24] S. Krek, Š. Arhar-Holdt, T. Erjavec, J. Čibej, A. Repar, P. Gantar, N. Ljubešić, I. Kosem, and K. Dobrovoljc, "Gigafida 2.0: the reference corpus of written standard Slovene," In: Proceedings of LREC 2020: Twelfth International Conference on Language Resources and Evaluation: May 11-16, pp. 3340-3345, Marseille, France, 2020.