FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Vojtěch Hudeček

Exploiting user's feedback to improve pronunciation of TTS systems

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University in Prague

Supervisor of the master thesis: doc. Ing. Zdeněk Žabokrtský, Ph.D.
Study programme: Informatics
Study branch: Artificial Intelligence

Prague 2017

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In .... date .... signature of the author

Title: Exploiting user's feedback to improve pronunciation of TTS systems

Author: Bc. Vojtěch Hudeček

Department: Name of the department

Supervisor: doc. Ing. Zdeněk Žabokrtský, Ph.D, Institute of Formal and Applied Linguistics

Abstract: Although spoken dialogue systems have greatly improved, they still cannot handle communications involving unknown topics and are very fragile. We will investigate methods that can improve spoken dialogue systems by correcting or even learn the pronunciation of unknown words. Thus we will provide better user experience, since for example mispronounced proper nouns are highly undesirable. Incorrect pronunciation is caused by imperfect phonetic representation, typically phonetic dictionary. We aim to detect incorrectly pronounced words by exploiting userâĂŹs feedback as well as using prior knowledge of the pronunciation and correct the transcriptions accordingly. Furthermore, the learned phonetic transcriptions can be used to improve speech recognition module by refining its models. Models used in speech recognition cannot handle words that are not in their vocabulary or have phonetic representation. Extracting those words from userâĂŹs utterances and adding them to the vocabulary should lead to a better overall performance.

Keywords: text-to-speech, automatic speech recognition, user's response, phonetic dictionary, machine learning, mel cepstral distortion

Dedication.

Contents

Introduction

1. Introduction

1.1 Introduction to the problematic

Voice control or communication is a common feature of many systems nowadays. Its applications ranges from simple one-word control commands to complex communication in spoken dialogue systems. In this work, we consider mainly these complex systems. For the sake of clarity, we now briefly describe setting of such system. It usually contains Automatic Speech Recognition (ASR) module, so the natural speech can be recognized and translated into words. The system then somehow derives an appropriate response, typically in the form of sentence written in natural language. This response can be displayed in the written form, however, it is more common to generate an audio with human voice reading the response. Although it is possible to use only a limited set of prerecorded utterances, it is desired to be able to read an arbitrary phrase. One reason is, that it may be difficult to read named entities and numerical values such as time and date. Also, the usage of variable utterances provides better user experience. Because of this, a Text-To-Speech (TTS) module is usually also part of dialogue systems. The purpose of this module is to transform a (generally arbitrary) written text to audio file containing the utterance read in natural voice. Modern TTS systems produce audio waveforms that sound quite naturally and the pronunciation is relatively good. Nevertheless, it have difficulties when it comes to so called Out-Of-Vocabulary (OOV) words. That is because the system is usually trained using certain set of words, typically from one language. But real applications often require to pronounce named entities or other language- or domain- specific words, that cannot be present during the training phase. This can cause situations, when words are mispronounced. Although it does not happen often, the negative effect can be quite strong, since it is inconvenient for the user when his or her name is pronounced with mistakes.

In this work, we aim to improve the TTS system pronunciation of desired words. To achieve this, we employ feedback gathered from the user. That means, we allow user to provide better example of pronunciation and thus we can correct ours. We are able to improve the TTS system by processing the obtained recording, deriving a phonetic transcription (i.e. pronunciation) and adding it to the TTS vocabulary. Moreover, the derived pronunciations can be used to improve the recognition ability of the ASR module. Also, we propose algorithm that can identify words, that are potentially difficult to pronounce without any prior language knowledge. More details regarding this issue can be found in respective sections. As it has been suggested, our method has got potentially very useful applications. It can be used to enlarge vocabulary of TTS or ASR systems both offline or on the fly using the live user's feedback. There exist several ways how to obtain such a feedback, however, this is not a subject of this work. Theoretically, the method can work with just one gold example, however, it is generally possible to ask user two or three times while not bothering him too much. (TODO: source) In dialogsample we provide basic example of simple dialogue, illustrating the process.

```
System:  Hello, /AANDRZHEZH/.
User:  You said it wrong, my name is /ONDRZHEI/.
System:  /ANDREY/, correct?
User:  No, it is /ONDRZHEI/.
System:  Oh, /ONDRZHEI/?
User:  That's right.
```

Figure 1.1: Sample dialogue illustrating the pronunciation correction. The transcriptions of the user's name are given in ARPABET**?**

Please note, that the dialogue policy is not part of this work and the way, how the feedback is obtained depends on the respective dialogue system.

1.2 Related Work

1.2.1 Grapheme-to-phoneme conversion

An automatic grapheme-to-phoneme conversion was first considered in the context of TTS applications. The input text needs to be converted to a sequence of phonemes which is then fed into a speech synthesizer. It is common in TTS systems that they first try to find the desired word in the dictionary and if it doesn't find it, it employs the grapheme-to-phoneme ($g2p$) module. A trivial approach is to employ a *dictionary look-up.* However, it cannot handle context and inherently covers only finite set of combinations. To overcome this limitations, the rule-based conversion was developed. Kaplan and Kay **?** formulate these rules in terms of finite-state automata. This system allows to greatly improve coverage. However the process of designing sufficient set of rules is difficult, mainly since it must capture irregularities. Because of this, a *data-driven* approach has to be introduced. Many machine learning techniques was explored, starting with Sejnowski and Rosenberg **?**. The approaches can be divided into three groups.

Techniques based on local similarities

The techniques presuppose an alignment of the training data between letters and phonemes or create such an alignment in a separate preprocessing step. The alignment is typically construed so that each alignment item comprises exactly one letter. Each slot is then classified (using its context) and a correct phoneme is predicted.

Pronunciation by analogy

These methods search for local similarities in the training lexicon. So it can be understood as a variation of the *nearest-neighbors* approach.

Probabilistic approaches

The problem can be looked at from a probabilistic point of view. One way to do this is to employ so called *joint sequence models* **?**. We use an open-source implementation of such a model in our work. This approach formalizes the task as follows:

$$\boldsymbol{\varphi}(\boldsymbol{g}) =_{\varphi' \in \Phi^*} p(\boldsymbol{g}, \boldsymbol{\varphi}') \tag{1.1}$$

where * denotes a Kleene star. In other words, for a given ortographic form $\boldsymbol{g} \in G^*$ we want to find the most likely pronunciation $\boldsymbol{\varphi} \in \Phi^*$, where G, Φ are ortographic and phonetic alphabets.

Generally, the problem of the wrong pronunciation in TTS is caused by a bad phonetic transcription. Traditional TTS systems are modular, one module's output is inputted into the next one. Because of this fact, the errors cumulate and thus the mistakes made by *g*2*p* cannot be repaired. So if we want to improve the pronunciation, we can try to improve the *g*2*p* as it is done in **?**. Authors in this work propose a method of exploiting a *g*2*p* trained on a language with a high number of available resources to create a *g*2*p* for language for which we do not have sufficient number of examples. This method relies on the existence of a conversion mapping between these two languages. Also it requires to do the conversion for every new language.

1.2.2 Learning pronunciation from spoken examples

. In theory, a model could be created which accepts an ortographic form together with language information and outputs a correct transcription. However, this information is typically not available. Also, if we want to learn a new pronunciation of just one word, it's more convenient to do it in a different way. Authors of **?** introduce method of deriving correct pronunciation for a word. They argue that in the spontaneous speech, the most frequent pronunciation doest not need to be the correct one. Thus they allow multiple way how to pronounce certain words. The pronunciations are obtained using n-best list of the modified recognizer. **?** and **?** improve this approach. Both these works modify the recognizer's training procedure and use an EM algorithm to estimate the parameters. They also introduce the concept of *graphones* - units consisting of pairs (*grapheme, phoneme*) and use it to create a language model for phoneme recognition.

1.3 Thesis overview

2. Title of the second chapter

2.1 Title of the first subchapter of the second chapter

2.2 Title of the second subchapter of the second chapter

Conclusion

@bookhuang2001spoken, title=Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, author=Huang, X. and Acero, A. and Hon, H.W., isbn=9780130226167, lccn=00050196, url=https://books.google.cz/books?id=reZQ year=2001, publisher=Prentice Hall PTR pp. 955-991. @bookegan1948test, title=Articulation Testing Methods, author=Egan, J., isbn=, lccn=, url=, year=1948, publisher=Laryngoscope ESPRIT SAM-UCL-042. @bookhoward92test, title=SOAP, Speech Output Assessment Package, author=Howard-Jones, P., isbn=, lccn=, url=, year=1992, publisher= @booknye74test, title=The Intelligibility of Synthetic Monosyllabic Words in Short, Syn- tactically Normal Sentences, author=Nye, P. and J. Gaitenby, isbn=, lccn=, url=, year=1974, publisher=Haskins Laboratories pp. 1047-1050 @bookacero99HMM, title=Formant Analysis and Synthesis Using Hidden Markov Models, author=Acero, A., isbn=, lccn=, url=, year=1999, publisher=Eurospeech @book, title=, author=, isbn=, lccn=, url=, year=, publisher= @inproceedingstokuda2015, title = Directly Modeling Speech Waveforms by Neural Networks for Statistical Parametric Speech Synthesis, author = Keiichi Tokuda and Heiga Zen, year = 2015, booktitle = Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages = 4215–4219

@booktaylor2009text, title=Text-to-Speech Synthesis, author=Taylor, P., isbn=9781139477 url=https://books.google.cz/books?id=T0O-NHZx7kIC, year=2009, publisher=Cambridge University Press

@ARTICLE2016arXiv160903499V, author = van den oord, A. and Dieleman, S. and Zen, H. and Simonyan, K. and Vinyals, O. and Graves, A. and Kalchbrenner, N. and Senior, A. and Kavukcuoglu, K., title = "WaveNet: A Generative Model for Raw Audio", journal = ArXiv e-prints, archivePrefix = "arXiv", eprint = 1609.03499, primaryClass = "cs.SD", keywords = Computer Science - Sound, Computer Science - Learning, year = 2016, month = sep, adsurl = http://adsabs.harvard.edu/abs/2016arXiv160903499V, adsnote = Provided by the SAO/NASA Astrophysics Data System

@inproceedingsharrison2009implementation, title=Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training., author=Harrison, Alissa M and Lo, Wai-Kit and Qian, Xiaojun and Meng, Helen, booktitle=SLaTE, pages=45–48, year=2009

@articlebisani2008joint, title=Joint-sequence models for grapheme-to-phoneme conversion, author=Bisani, Maximilian and Ney, Hermann, journal=Speech Communication, volume=50, number=5, pages=434–451, year=2008, publisher=Elsevier

@inproceedingsratanamahatana2004everything, title=Everything you know about dynamic time warping is wrong, author=Ratanamahatana, Chotirat Ann and Keogh, Eamonn, booktitle=Third Workshop on Mining Temporal and Sequential Data, year=2004, organization=Citeseer

@inproceedingskubichek1993mel, title=Mel-cepstral distance measure for objective speech quality assessment, author=Kubichek, R, booktitle=Communications, Computers and Signal Processing, 1993., IEEE Pacific Rim Conference on, volume=1, pages=125–128, year=1993, organization=IEEE @articlefrey2007clustering, title=Clustering by passing messages between data points, author=Frey, Brendan J and Dueck, Delbert, journal=science, volume=315, number=5814, pages=972–976, year=2007, publisher=American Association for the Advancement of Science

@articlenavarro2001guided, title=A guided tour to approximate string matching, author=Navarro, Gonzalo, journal=ACM computing surveys (CSUR), volume=33, number=1, pages=31–88, year=2001, publisher=ACM

@inproceedingsmolau2001computing, title=Computing mel-frequency cepstral coefficients on the power spectrum, author=Molau, Sirko and Pitz, Michael and Schluter, Ralf and Ney, Hermann, booktitle=Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on, volume=1, pages=73–76, year=2001, organization=IEEE

@articletaylor1997ssml, title=SSML: A speech synthesis markup language, author=Taylor, Paul and Isard, Amy, journal=Speech communication, volume=21, number=1-2, pages=123–133, year=1997, publisher=Elsevier

@miscArpabet, title = Arpabet overview, howpublished = `https://nlp.stanford.edu/courses/lsa352/arpabet.html`, note = Accessed: 2017-04-16

@articlewang2017tacotron, title=Tacotron: A Fully End-to-End Text-To-Speech Synthesis Model, author=Wang, Yuxuan and Skerry-Ryan, RJ and Stanton, Daisy and Wu, Yonghui and Weiss, Ron J and Jaitly, Navdeep and Yang, Zongheng and Xiao, Ying and Chen, Zhifeng and Bengio, Samy and others, journal=arXiv preprint arXiv:1703.10135, year=2017

@articlevan2016wavenet, title=Wavenet: A generative model for raw audio, author=van den Oord, Aäron and Dieleman, Sander and Zen, Heiga and Simonyan, Karen and Vinyals, Oriol and Graves, Alex and Kalchbrenner, Nal and Senior, Andrew and Kavukcuoglu, Koray, journal=CoRR abs/1609.03499, year=2016

@articleyao2015sequence, title=Sequence-to-sequence neural net models for grapheme-to-phoneme conversion, author=Yao, Kaisheng and Zweig, Geoffrey, journal=arXiv preprint arXiv:1506.00196, year=2015

@inproceedingsschlippe2012grapheme, title=Grapheme-to-phoneme model generation for Indo-European languages, author=Schlippe, Tim and Ochs, Sebastian and Schultz, Tanja, booktitle=Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pages=4801–4804, year=2012, organization=IEEE

@articlekaplan1994regular, title=Regular models of phonological rule systems, author=Kaplan, Ronald M and Kay, Martin, journal=Computational linguistics, volume=20, number=3, pages=331–378, year=1994, publisher=MIT Press

@booksejnowski1988nettalk, title=NETtalk: A parallel network that learns to read aloud, author=Sejnowski, Terrence J and Rosenberg, Charles R, year=1988, publisher=MIT Press

@articlebisani2008joint, title=Joint-sequence models for grapheme-to-phoneme conversion, author=Bisani, Maximilian and Ney, Hermann, journal=Speech communication, volume=50, number=5, pages=434–451, year=2008, publisher=Elsevier

@inproceedingsderi2016grapheme, title=Grapheme-to-phoneme models for (almost) any language, author=Deri, Aliya and Knight, Kevin, booktitle=Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, volume=1, pages=399–408, year=2016

@inproceedingsslobada1996dictionary, title=Dictionary learning for spontaneous speech recognition, author=Slobada, Tilo and Waibel, Alex, booktitle=Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, volume=4, pages=2328–2331, year=1996, organization=IEEE

@articlemcgraw2013learning, title=Learning lexicons from speech using a pronunciation mixture model, author=McGraw, Ian and Badr, Ibrahim and Glass, James R, journal=IEEE Transactions on Audio, Speech, and Language Processing, volume=21, number=2, pages=357–366, year=2013, publisher=IEEE

@inproceedingsreddy2011learning, title=Learning from Mistakes: Expanding Pronunciation Lexicons Using Word Recognition Errors., author=Reddy, Sravana and Gouvêa, Evandro B, booktitle=INTERSPEECH, pages=533–536, year=2011

@inproceedingskubichek1993mel, title=Mel-cepstral distance measure for objective speech quality assessment, author=Kubichek, R, booktitle=Communications, Computers and Signal Processing, 1993., IEEE Pacific Rim Conference on, volume=1, pages=125–128, year=1993, organization=IEEE

@inproceedingsvan2008autonomata, title=The AUTONOMATA Spoken Names Corpus., author=Van Den Heuvel, Henk and Martens, Jean-Pierre and D'hoore, Bart and D'hanens, Kristof and Konings, Nanneke, booktitle=LREC, year=2008

@articlesalvador2007toward, title=Toward accurate dynamic time warping in linear time and space, author=Salvador, Stan and Chan, Philip, journal=Intelligent Data Analysis, volume=11, number=5, pages=561–580, year=2007, publisher=IOS Press

@inproceedingsng2002spectral, title=On spectral clustering: Analysis and an algorithm, author=Ng, Andrew Y and Jordan, Michael I and Weiss, Yair, booktitle=Advances in neural information processing systems, pages=849–856, year=2002

List of Figures

List of Tables

List of Abbreviations

Attachments