Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 46 (2017) 1–35

www.elsevier.com/locate/csl

Anomaly-based annotation error detection in speech-synthesis corpora☆

Jindřich Matoušek*, Daniel Tihelka

Department of Cybernetics, New Technology for the Information Society (NTIS), Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, Pilsen 306 14, Czech Republic

Received 26 September 2016; Accepted 11 April 2017

Abstract

We investigate the problem of automatic detection of annotation errors in single-speaker read-speech corpora used for speech synthesis. For the purpose of annotation error detection, we adopt an anomaly detection framework in which correctly annotated words are considered as normal examples on which the detection methods are trained. Misannotated words are then taken as anomalous examples which do not conform to normal patterns of the trained detection models. We propose and evaluate several anomaly detection models – Gaussian distribution based detectors, Grubbs' test based detector, and one-class support vector machine based detector. Word-level feature sets including basic features derived from forced alignment and various acoustic, spectral, phonetic, and positional features are examined to find an optimal set of features for each anomaly detector. The results with $F1$ score being almost 89% show that anomaly detection could help detecting annotation errors in read-speech corpora for speech synthesis. Furthermore, dimensionality reduction techniques are also examined to automatically reduce the number of features used to describe the annotated words. We show that the automatically reduced feature sets achieve statistically similar results as the hand-crafted feature sets. We also conducted additional experiments to investigate both robustness of the proposed anomaly detection framework with respect to particular data sets used for development and evaluation and the influence of the number of examples needed for anomaly detection. We show that a reasonably good detection performance could be reached with using significantly fewer examples during the detector development phase. We also propose a concept of a voting detector – a combination of anomaly detectors in which each "single" detector "votes" on whether or not a testing word is annotated correctly, and the final decision is then made by aggregating the votes. Our results show that the voting detector has a potential to overcome each of the single anomaly detectors. Furthermore, we compare the proposed anomaly detection framework to a classification-based approach (which, unlike anomaly detection, needs to use anomalous examples during training) and we show that both approaches lead to statistically comparable results when all available anomalous examples are utilized during detector/classifier development. However, when a smaller number of anomalous examples are used, the proposed anomaly detection framework clearly outperforms the classification-based approach. A final listening test showed the effectiveness of the proposed anomaly-based annotation error detection for improving the quality of synthetic speech.

Keywords: Annotation error detection; Anomaly detection; Read speech corpora; Speech synthesis

☆ This paper has been recommended for acceptance by Simon King.

* Corresponding author.
E-mail address: jmatouse@kky.zcu.cz (J. Matoušek).

http://dx.doi.org/10.1016/j.csl.2017.04.007
0885-2308/

1. Introduction

Nowadays, the most successful speech processing methods utilize very large speech corpora. The advantage of the so-called *corpus-based methods* is that large corpora can capture the inherent variability of human speech. In addition to speech recordings themselves, the corpora also contain some annotation data which describe what was actually pronounced in each recording. Though many annotation levels may exist (such as phonetic, syntactic, morphological, etc.), a *word-level annotation* (also known as *orthographic annotation*) usually constitutes the basis from which the other levels are derived.

Although some attempts were made to annotate corpora automatically or semi-automatically (see, e.g., Cox et al., 1998; Meinedo and Neto, 2003; Adell et al., 2006; Hazen, 2006; Tachibana et al., 2007; Aylett et al., 2009; Boeffard et al., 2012), the automation is still error-prone. On the other hand, manual annotation is a time-consuming and costly process, and, in addition, even human annotators still make errors. Typical annotation errors are missing or extra words, swapped, mispronounced or otherwise misannotated words (Matoušek and Romportl, 2007). A more detailed analysis of annotation errors will be further provided in Section 2.1.

It is obvious that word-level annotation of speech data is still one of the most important processes for many speech-processing tasks. It is also evident there is a need to have the word-level annotation as accurate as possible because any discrepancy between speech signal and its linguistic representation is critical, especially in methods which use the speech signal directly (such as concatenative speech synthesis). Although the annotation error detection methods described in this paper could be used for the detection of annotation errors in virtually any corpus, we will further focus on read-speech corpora typical for speech synthesis. Such corpora are somewhat specific as they typically contain the speech of a single speaker (often a professional voice talent) with a very consistent style of speaking in terms of consistent speech tempo, pitch, timbre etc.

In text-to-speech (TTS) synthesis, both methods currently widely used – *unit selection* and *statistical parametric speech synthesis* (SPSS) – do suffer from imperfect word-level annotation. In the case of unit selection (a signal-based concatenative synthesis approach) (Hunt and Black, 1996), the resulting synthetic speech is made up of directly concatenating natural signals of speech segments. The negative impact of wrong word-level annotation is perhaps more evident in this speech synthesis method because any mismatch between speech signal (which is believed to have linguistic properties derived from the corresponding word-level annotation) and its annotation may inherently result in serious audible glitches in synthetic speech – synthesized speech could be unintelligible, or even other speech than expected may be synthesized (Matoušek et al., 2012). The impact of annotation errors on the quality of unit-selection based synthetic speech is further analyzed in Section 2.2. In the case of SPSS (a model-based generative speech synthesis approach) (Zen et al., 2009), synthetic speech is generated from relevant speech-related parameters modeled by statistical models. In this approach, the wrong annotation causes a discrepancy between linguistic and acoustic parts of the SPSS system – acoustic models are then simply trained on inappropriate data. Although the statistical framework used within SPSS is able to filter out these discrepancies to a certain extent, it was shown that mismatch between the linguistic and acoustic parts could grossly degrade the acoustic models, and that a correction of these mismatches (even on the level of pronunciation variants of a word) can improve the synthesis quality of synthetic speech (Dall et al., 2016).

Generally, word-level annotation error detection could be viewed as a part of the "checking and tidying data" steps commonly used in many speech-processing tasks. These steps are considered as an important part of data preprocessing and are extensively used in the field of data science, for instance in machine learning pipelines. The importance of data checking and tidying may increase even more in today's world of big data because given the large amount of data manual checking is often not feasible.

Word-level annotation errors are often further manifested as phonetic transcription and segmentation errors. As the automatic phonetic segmentation accuracy has attracted researchers for many years, two main segmentation schemes were proposed, with the *hidden Markov models* (HMM) based forced alignment framework (Donovan and Woodland, 1999) (including a number of refinements – see e.g., Toledano et al., 2003; Park and Kim, 2007; Lin and Jang, 2007; Matoušek and Romportl, 2008; Rendel et al., 2012) being generally preferred over the *dynamic time warping* (DTW) framework (Horák, 2002; Malfrere et al., 2003; Paulo and Oliveira, 2003). On the other hand, the origin of gross segmentation errors and a way of fixing them has not been researched so much. Instead, erroneous segments, if detected, are usually discarded, and other segments are selected in unit-selection speech synthesis. As discussed by Taylor (2009), the chase for ideal phonetic segmentation may not be so important; automatic