Available online at www.sciencedirect.com
ScienceDirect
Computer Speech & Language xxx (2017) xxx-xxx
www.elsevier.com/locate/csl

On the use of acoustic features for automatic disambiguation of homophones in spontaneous German ☆

Barbara Schuppler*, Tobias Schrank

Signal Processing and Speech Communication Laboratory, Graz University of Technology, Inffeldgasse 16c, 8010 Graz, Austria

Received 27 January 2017; received in revised form 14 October 2017; accepted 28 December 2017
Available online xxx

Abstract

Homophones pose serious issues for automatic speech recognition (ASR) as they have the same pronunciation but different meanings or spellings. Homophone disambiguation is usually done within a stochastic language model or by an analysis of the homophonous word's context, similarly to word sense disambiguation. Whereas this method reaches good results in read speech, it fails in conversational, spontaneous speech, where utterances are often short, contain disfluencies and/or are realized syntactically incomplete. Phonetic studies, however, have shown that words that are homophonous in read speech often differ in their phonetic detail in spontaneous speech. Whereas humans use phonetic detail to disambiguate homophones, this linguistic information is usually not explicitly incorporated into ASR systems. In this paper, we show that phonetic detail can be used to automatically disambiguate homophones using the example of German pronouns. Using 3179 homophonous tokens from a corpus of spontaneous German and a set of acoustic features, we trained a random forest model. Our results show that homophones can be disambiguated reasonably well using acoustic features (74% F_1, 92% accuracy). In particular, this model is able to outperform a model based on lexical context (48% F_1, 89% accuracy). This paper is of relevance for speech technologists and linguists: amodule using phonetic detail similar to the presented model is suitable to be integrated in ASR systems in order to improve recognition. An approach similar to the work here that combines the automatic extraction of acoustic features with statistical analysis is suitable to be integrated in phonetic analysis aiming at finding out more about the contribution and interplay of acoustic features for functional categories.

Keywords: Homophone disambiguation; Automatic speech recognition; Phonetic detail; Spontaneous speech; Random forests

1. Introduction

Homophones and near-homophones pose serious difficulties for automatic speech recognition (ASR) (Goldwater et al., 2008). They have the same or – as in the case of near-homophones – a similar pronunciation but different meanings or spellings. If an ASR system needs to deal with a homophonous word, it needs to decide which lexeme underlies this word in order to perform well. This process is called homophone disambiguation. Homophone

☆ This paper has been recommended for acceptance by Roger Moore.
* Corresponding author.
E-mail address: b.schuppler@tugraz.at (B. Schuppler), tobias.schrank@tugraz.at (T. Schrank).

http://dx.doi.org/10.1016/j.csl.2017.12.011
0885-2308/ 2017 Published by Elsevier Ltd.

disambiguation is usually done within a stochastic language model (Lee, 2003) or by an analysis of the homophonous word's context, similarly to word sense disambiguation (Béchet et al., 1999; Jurafsky and Martin, 2009). While this context-based form of homophone disambiguation is often successful, it is not for homophones that share similar syntactic contexts, so-called doubly confusable pairs (e.g., *they asked him* and *they ask him*) (Goldwater et al., 2010). Whereas it has been suggested to exploit more syntactic and discursive information to distinguish between members of doubly confusable pairs (Goldwater et al., 2010), we propose to exploit acoustic cues. This proposal is motivated by two reasons: (1) Performing homophone disambiguation by inspection of the homophone's lexical context may lead to increased word error rate due to recognition errors in the homophone's context as shown in the work by Béchet et al. (1999). This is especially true in spontaneous speech that contains breaks, repairs and similar discontinuities. (2) A number of phonetic studies highlighted differences in phonetic detail between homophones, especially for homophones produced in spontaneous speech (e.g., Ward, 2004; Gahl, 2008; Nemoto et al., 2008; Niebuhr and Kohler, 2011; Samlowski et al., 2013; Volín et al., 2014). To our knowledge, however, such differences in phonetic detail have not yet been used for homophone disambiguation in an ASR system.

In the last decade, there was a growing interest in studying the predictors for pronunciation variation (see Section 1.1). Besides the well studied predictors such as segmental context, word frequency and phrase position, we hypothesize that the realization of a word also depends on its morphosyntactic attributes. It has also been shown that differences in word duration can aid in learning syntactic structures (Pate and Goldwater, 2013). We, however, propose to look at a more constrained line of research that is directly applicable to ASR. If morphosyntactic information is directly encoded in the speech signal, then many homophones can be disambiguated using acoustic features alone. As there is generally more variation in spontaneous speech (e.g., Ostendorf et al., 2003; Nakamura et al., 2008), we expect these differences to be particularly pronounced in spontaneous speech. Moreover, our research is also particularly relevant for spontaneous speech for another reason: Due to the high amount of reduction in spontaneous speech, there are more phonologically homophonous tokens than in read speech (Niebuhr and Kohler, 2011).

In order to test our hypothesis that homophones can be disambiguated acoustically, we analyzed the German words ⟨der⟩ [deːɐ], ⟨die⟩ [diː], ⟨das⟩ [das] and their inflections ⟨des⟩ [dɛs], ⟨dem⟩ [deːm], ⟨den⟩ [deːn]. Each of these word forms take either the function of determiner (DET), relative pronoun (REL) or demonstrative pronoun (DEM). All of these can surface in similar contexts[1]:

(1) *der* 13. November passt
DET ADJ NOUN VERB
November 13th is fine
(2) der Freitag *der* nach Ostern kommt passt
DET NOUN REL ADP NOUN VERB VERB
the Friday after Easter is fine
(3) Freitag *der* passt
NOUN DEM VERB
Friday, that is fine

All grammatical functions of a word form share the same phonological form. Despite this, significant acoustic differences between different functions of the same word forms could be found in a controlled reading task (Samlowski et al., 2013). This paper aims at making these findings usable for ASR in spontaneous speech. This is especially relevant, as these function words occur frequently in spontaneous conversation (e.g., 68% of all utterances in the Kiel Corpus of Spontaneous Speech (Kohler et al., 1995) contain at least one instance of the mentioned word forms). What is more, our approach of using acoustic features for homophone disambiguation can be applied to other types of homophones.

In this paper, we automatically extracted acoustic features from 3179 realizations of homophonous word forms. We analyzed these acoustic features and searched for systematic differences between the realizations of the same word form. We then used this information to automatically disambiguate homophones by means of random forests. In order to learn more about the variation of homophonous structures we discuss the relevance of each feature class.

[1] In this paper, we used a combination of part-of-speech (POS) tags as developed in Petrov et al. (2012) and category labels as developed in Bickel et al. (2008) (e.g., SG).

Please cite this article as: B. Schuppler, T. Schrank, On the use of acoustic features for automatic disambiguation of homophones in spontaneous German, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.12.011