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

Barbara Schuppler[a,*], Tobias Schrank[a]

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

Abstract

Homophones pose serious issues for automatic speech recognition (ASR). In order to deliver high quality ASR output, homophones need to be disambiguated. Homophone disambiguation is usually done by analysing the homophonic word's context. 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. From phonetic studies, however, we learned that words which are homophonic in read speech often differ in their phonetic detail in spontaneous speech. Whereas humans use phonetic detail to disambiguate homophones, ASR systems usually ignore it. In this paper, we show that phonetic detail can be used to automatically disambiguate homophones. For our experiments, we use 3146 homophonic tokens from a corpus of spontaneous German. We collect a set of acoustic features and train a random forest model. Our results show that homophones can be disambiguated reasonably well using acoustic features (71% F_1, 92% accuracy). In particular, this model is able to outperform a model based on lexical context (48% F_1, 89% accuracy). A module using phonetic detail similar to our model is suitable to be integrated in ASR systems in order to improve word recognition.

Keywords: homophone disambiguation, automatic speech recognition, phonetic detail, spontaneous speech, random forests

*Corresponding author

Email addresses: b.schuppler@tugraz.at (Barbara Schuppler), tobias.schrank@tugraz.at (Tobias Schrank)

1. Introduction

Homophones and near-homophones pose serious issues for automatic speech recognition (ASR) (Goldwater et al., 2008). If an ASR system encounters a homophonic word, it needs to decide which lexeme underlies this word in order to deliver high quality output. This process is called homophone disambiguation. Homophone disambiguation is usually done within a stochastic language model (Lee, 2003) or by an analysis of the homophonic 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 (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) Strong syntactic constraints worsen word error rate (Béchet et al., 1999). This is especially true in spontaneous speech which contains breaks, repairs and similar discontinuities. (2) A number of phonetic studies highlighted differences in phonetic detail between homophones (e.g., Ward (2004); Gahl (2008); Rena Nemoto and Adda-Decker (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. In this regard, it has already 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)), 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 homophonic tokens than in read speech (Niebuhr and Kohler, 2011).

In order to test our hypothesis that homophones can be disambiguated acoustically, we analyze the German word forms ⟨der⟩ [deːɐ̯], ⟨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) Freitag *der* 13. passt
NOUN DET ADJ VERB
Friday 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 the articles, and the demonstrative and relative pronouns 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 these word forms). Furthermore, our approach of using acoustic features for homophone disambiguation can be applied to other types of homophones.

In this paper, we automatically extract acoustic features from 3184 realizations of homophonic word forms. We analyze these acoustic features and search for systematic differences between the realizations of the same word form. We then use this information to automatically disambiguate homophones by means of random forests. In order to learn more about the variation of homophonic structures and about the relevance of each feature class, we further analyze our data by means of mixed effects logistic regression

[1]In this paper, we use 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).

models. Furthermore, we also propose a way to integrate this information in current ASR system designs and thus improve word recognition.

1.1. Phonetic detail in speech production and perception

Phonetic production experiments and corpus studies have shown that the realization of the same word differs in dependence of several factors: One of the best understood type of factors are the connected speech factors (e.g., speech rate (Jurafsky et al., 1998), speaking style (Ernestus et al., 2015), segmental context (Schuppler et al., 2012)). Other parameters which are known to affect the pronunciation of a word are the speaker's social identity (Drager, 2011), word frequency (Gahl, 2008), bigram frequency (e.g., Schuppler et al. (2012); Torreira and Ernestus (2009)), word predictability (Jurafsky et al., 2002), and word identity (Pierrehumbert, 2002). Furthermore, also syntactic structure (Pate and Goldwater, 2013) and morphological properties have been shown to affect the degree of reduction of a word. For English, Baker et al. (2007) found that the prefix *mis* has longer duration in words where it functions as a productive morpheme (e.g., *mistimes*) than in words where it is a non-productive pseudo morpheme (e.g., *mistakes*). Similarly, Schuppler et al. (2012) found that homophoneous word pairs differ in their phonetic detail depending on their morphological properties: For instance, the word final /t/ in the Dutch word *vind* '(I) find' with the canonical pronunciation [vɪnt] is more likely to be reduced than the /t/ in the homophoneous word *vindt* '(he) finds' where the final /t/ also carries a morphological function. Finally, Samlowski et al. (2013) found in German read speech that homophoneous demonstrative pronouns, relative pronouns and definite articles differ in duration, prominence and spectral characteristics. In this paper, we aim at investigating whether also in German spontaneous speech homophoneous demonstrative pronouns, relative pronouns and definite articles differ with respect to their phonetic detail. In our set of features, we include temporal and spectral features which reflect mentioned vowel and consonant reductions and deletions.

Furthermore, the different pragmatic (Plug, 2006) and interactional functions (Local, 2003) of words have been shown to differ with respect to their phonetic detail. A quantitative approach to studying the relationship between phonetic detail and communicative function and/or meaning has been presented by Volín et al. (2014). In their study, they used a set of 36 prosodic features to automatically classify eight functional categories of the Czech affirmation particle *jasně*. They compared three different statistical approaches

for classification: discriminant analysis, regression trees and artificial neural networks and interpreted the outcome linguistically. They conclude that each of the classification outcomes significantly reflect that the prosodic variation is linked systematically with the different functional categories of *jasně* and that their methods are suitable to learn about the relative importance and the interplay of the individual predictors for the different categories. Also in this paper, we will use statistical modeling techniques to study the interplay of acoustic features (a set of 189 features) and to investigate whether also grammatical categories (determiner, relative pronoun or demonstrative pronoun) can be distinguished automatically (in contrast to classifying functional categoris as in Volín et al. (2014)). Instead of using discriminant analysis and artificial neural networks, we prefer to use random forests (L., 2001), a relatively novel statistical approach which serves well to discover context-dependent relationships and which deals well with highly correlated variables (see Section 3.2 for more details on the method). In addition, we will use mixed effects logistic regression (Jaeger, 2008) in order to study the details of the interplay of certain factors.

Based on the assumption that humans use phonetic detail particularly extensively when disambiguating homophones (Hawkins and Smith, 2001), a series of experiments revealed information about which type of phonetic detail humans actually use in speech perception. It has been shown that the perceived meaning of junctural minimal pairs (e.g., *So he diced them* vs. *So hed iced them* is effected by the voice of individual speakers. Smith and Hawkins (2012) found in a production experiment that the speakers vary significantly as in how they use certain acoustic features to distinguish such junctural minimal pairs. A subsequent perception experiment designed to test the intelligibility of those junctural minimal pairs in noise revealed that when listeners hear tokens spoken by the same voice as in the familiarization period they perform better with regard to the identification of words and the syllable boundaries than when they hear tokens spoken by a different voice. Not specific for homophones, but for the perception of word meaning in general, Nzgaard et al. (2009) showed in a production and perception experiment revealed that speech contains prosodic characteristics specific for word meaning and that listeners also use these cues to disambiguate meanings. They conclude that their study provides evidence for the existence of prosodic correlates to word meaning. Finally, Nzgaard and Lunders (2002) have shown that the emotional tone of voice is relevant for the processing of lexically ambiguous words and affects the selection of the word meaning.

Whereas Nzgaard and Lunders (2002) showed the effect of tone of voice for native speakers, Hanulikovà and Haustein (2016) found effects of emotional prosody on the processing of lexically ambiguous words also for non-native speakers. In this paper, we will not present any perception experiments. We will, however, integrate acoustic features for the automatic disambiguation of the homophones which are related to the here presented findings from speech perception (e.g., we use prosodic and spectral features which also showed good results in emotion recognition (Schuller et al., 2007)).

1.2. Phonetic detail and ASR for spontaneous speech

In comparison to read speech, spontaneous speech poses serious problems for automatic speech recognition (Nakamura et al., 2008, p.172): "Spontaneous speech can be characterized by accelerated speaking rate, sloppy pronunciation, filled pauses, repairs, hesitations, repetitions, partial words, and disfluencies." In order to be able to deal with these difficulties, Nakamura et al. (2008) suggests to " (a) analyzing quantitative differences between spontaneous and read speech, and (b) clarifying the reason why the recognition performance for spontaneous speech is low [p.172]." A series of studies has focused on the misrecognitions of ASR systems and finding the factors which cause the errors[2]. In such an error analysis of ASR systems for English and French spontaneous speech, Adda-Decker and Lamel (2005) found that male speakers are harder to recognize than female speakers. Their detailed phonetic analysis revealed that the speech of men contains more temporal segment reductions. Bell et al. (2002) also found that women are more likely to use fuller forms, even when controlling for speech rate. Another evidence for reduced words being difficult to recognize comes from Nakamura et al. (2008), who found that the reduction of the MFCC space of vowels and consonants directly contributes to the decrease of speech recognition performance for spontaneous speech in comparison to read speech. Moreover, these reductions are especially relevant for fast speech. Siegler and Stern (1995) observed among others that ASR performance drops when speech rate increases. Also Shinozaki and Furui (2001) provide a detailed error analysis for automatic speech recognition of conversational Japanese. They found higher word error rates for low probability words and for short words (similar as Goldwater et al. (2008)) as well as for utterances produced at higher

[2]For an overview of the role of speech variability see Benzeghiba et al. (2007)

speaking rate. Finally, a fairly large percentage of misrecognized words in ASR are connected to specific words. In other words, there are words that are harder to recognize than others. These words which are harder to recognize are mostly confused with homophones or near-homophones Goldwater et al. (2008), which is exactly the focus of this paper.

Rena Nemoto and Adda-Decker (2008) dealt with homophone disambiguation of the French words *et* 'and' and *est* 'to be' in spontaneous broadcast speech. Motivated by the high word error rate (WER) for these words (25% for *et* and 20% of *est*), they compared their ASR results with the performance of humans in a perception test. The participants listened or read the homophones in a 7-gram context (3-gram left, 3-gram right). In general, they concluded that humans do five times better than the ASR system, but that the conditions which were especially difficult for ASR also posed problems for the human listeners. Furthermore, their results suggest that listeners make use of prosodic/acoustic information especially when homophones occur in ambiguous syntactic structures. In order to find those acoustic features that aid homophone disambiguation, they tested 25 different classification algorithms with a set of 41 acoustic features (e.g., duration, f0, pauses) and reached accuracies of the best performing algorithm of 77% for *et* and 78% for *est*. Rena Nemoto and Adda-Decker (2008) subsequently suggested that the use of acoustic information might be useful not only for humans but also for an ASR system. They did, however, not provide information on how to implement that type of information into an ASR system nor did they provide performance comparisons with the traditional context-based approach. This paper investigates whether our large set of acoustic features (189) performs well on homophone disambiguation and whether the performance of acoustic features compares favorably to lexical features.

1.3. Our approach

To improve homophone disambiguation, we slightly change the way an ASR system uses its two main information sources: the acoustic signal and the language model[3]. Within the probabilistic paradigm, ASR in its most basic form is the act of searching for the most probable word sequence $W = w_1, \ldots, w_n$ given acoustic observations $x = x_1, \ldots, x_T$:

$$W^* = \arg\max_W p(W|x) \tag{1}$$

[3]We here ignore the pronunciation model as it is irrelevant for the current discussion

or using Bayes rule

$$W^* = \arg\max_W \frac{p(x|W)p(W)}{p(x)}, \tag{2}$$

in which $p(x)$ does not affect the maximization and thus can be ignored. Homophone disambiguation is usually performed on the level of the language model $p(W)$. For large vocabulary ASR this is generally a trigram model:

$$p(W) = p(w_1)p(w_2|w_1)\prod_{i=3}^{n} p(w_i|w_{i-1}, w_{i-2}). \tag{3}$$

In speech recognition, the acoustic model discovers phones and discards noise Jurafsky and Martin (2009). What is considered noise in this paradigm, however, is not only noise but contains relevant information: phonetic detail. Humans use phonetic detail particularly extensively when disambiguating homophones Hawkins and Smith (2001). This is to say that phonologically homophonic structures are potentially phonetically distinct. In order to show this in a practical setting, we disambiguate homophones in the following way: After word recognition – in our experiments this is replaced by manual transcriptions done by human transcribers – we treat homophonic words as separate tokens. For instance, ⟨der⟩ is now split into one token der_{DET}, one token der_{REL} and another token der_{DEM}.We then extract acoustic features designed to represented phonetic detail. We use these features to train a learner which rescores the ASR output. Through this rescoring step, the language model now also depends on acoustic features:

$$p(w') = p(w)p(w|\theta), \tag{4}$$

in which θ represents features of phonetic detail. Our approach leaves the fundamental architecture of the ASR system unchanged. Through its simplicity, the overhead can be kept low as rescoring is necessary only for homophonic words.

2. Material and annotation

Our analysis is based on spontaneous speech data from the Kiel Corpus of Spontaneous Speech (Kohler et al., 1995). We chose this speech material as it comes with detailed transcriptions at the orthographic, segmental and

Table 1: Absolute frequencies of category labels per word form.

	DET	REL	DEM	total
⟨der⟩	472	4	17	493
⟨die⟩	276	28	20	324
⟨das⟩	153	3	1148	1304
⟨des⟩	38	0	0	38
⟨dem⟩	275	0	2	277
⟨den⟩	714	5	29	748
total	1928	40	1216	3184

supra-segmental level. Moreover, there are already detailed phonetic studies available which are particularly focused on acoustic characteristics of reductions present in the speech material (e.g., Kohler et al. (1996); Kohler and Rodgers (2001); Wesener (1999)). The Kiel Corpus of Spontaneous Speech contains 126 conversations from 18 speaker pairs, each conducting 7 dialogues. These 36 speakers produced a total of 4721 utterances in 2061 turns which equals to 42945 tokens. Of these, ⟨der, die, das, des, dem, den⟩ account for 3184 tokens. Of all 2061 turns, 1406 turns (68%) contain at least one instance of ⟨der, die, das, des, dem, den⟩ and are therefore likely to contain recognition errors due to the multi-functionality of these word forms.

We annotated all target words with information regarding their function: determiner, relative pronoun or demonstrative pronoun. For the distribution of annotation labels see Table 1. We also annotated all target words with information regarding their gender (feminine, masculine, neuter), number (singular, plural) and case (nominative, genitive, dative, accusative). However, this information is not used in the present analysis due the very low number of observations in most of the resulting classes (e.g., there are only two instances of ⟨dem⟩ as demonstrative pronoun). The 38 instances of ⟨des⟩ had to be excluded completely from the following analysis because they surface only in the function of a determiner.

It also has to be noted that there is no unique relation mapping word forms to lexemes for ⟨der, die, das⟩. No word form corresponds to only one inflected form. With the exception of ⟨das⟩, all word forms can also be the surface form of at least two different lexemes. See Table 2 for an overview. These words share both acoustic and syntactic properties are also

Table 2: **Word form–lexeme relation:** The lexemes at the top take the respective word form on the left if inflected for case and number as shown in the fields.

	lexeme		
word form	der	die	das
⟨der⟩	1 SG, 2 PL	2 3 SG, 2 PL	2 PL
⟨die⟩	1 4 PL	1 4 SG, 1 4 PL	1 4 PL
⟨das⟩			1 4 SG
⟨des⟩	2 SG		2 SG
⟨dem⟩	3 SG		3 SG
⟨den⟩	4 SG, 3 PL	3 PL	3 PL

called doubly confusable pairs Goldwater et al. (2010). The examples 4 and 5 illustrate this point with a pair of German sentences that are identical sequences of words.

(4) Hans, der Floh, hatte ein gutes Leben.
John, the flea, had a good life.

(5) Hans, der floh, hatte ein gutes Leben.
John, who fled, had a good life.

If these two sentences are the output of an ASR system and thus capitalization is absent, they cannot be distinguished without further contextual information. However, only in example 4, ⟨der⟩ (canonical realization [dɛɐ̯]) – the masculine singular nominative determiner – can be reduced to the forms [dɐ̯], [də] or [d]. In example 5, ⟨der⟩ – the masculine singular nominative relative pronoun– can take neither of these reduced forms. It is always pronounced as [dɛɐ̯]. This difference is not due to differences in syntactic structure but is lexically fixed. With our acoustic-based approach to homophone disambiguation, we aim at making use of these differences with respect to phonetic detail.

We do not take any measures to assure that the target words occur in the same prosodic context. We do, however, include their position in the clause as a feature. This is particularly important as some positions are known to lead to articulatory strengthening (Goldwater et al., 2010). All target words tend to occur phrase-initially. Yet, demonstrative pronouns may behave differently as they frequently receive contrastive stress in the present speech data. We do not control other factors even though they are

known to influence the realization of words as well (e.g., syntactic factors (Jurafsky et al., 1998), segmental context: for instance segments tend to be less reduced when the following word begins with a vowel than when it begins with a consonant (Jurafsky et al., 1998), the speaker's dialect).

3. Method

3.1. Feature grouping and feature extraction

For feature extraction, we use Praat Boersma and Weenink (2013), and the R packages tuneR Ligges et al. (2014) and seewave Sueur et al. (2008). We collect the following 189 features from each word and group them into five feature classes:

1. **Temporal** (10 features): Word duration, segment durations, onset duration, nucleus duration, coda duration. **Segmental deletions** are indirectly encoded as segments having 0 duration.
2. **Fundamental frequency** (19 features): Static descriptors (e.g., extremes, position of extremes, mean, higher order moments, coefficients of linear regression).
3. **Intensity** (19 features): Static descriptors (e.g., extremes, position of extremes, mean, higher order moments, coefficients of linear regression).
4. **Spectral** (138 features): Coefficients of parabolic regression of formants 1–3, perceptual linear predictives (PLP), long term average spectrum (LTAS).
5. **Other** (3 features): Speaker gender (male or female), speaker identity, position of word in the clause.

All feature values are normalized by speaker using the respective z-scores. The values of temporal features are also normalized by speech rate. Speech rate is defined as clause duration divided by the number of words contained in this clause. No feature values are manually corrected.

3.2. Feature selection and classification with random forests

Since the main aim of this paper is to find phonetic characteristics which are useful for improving ASR, we do not focus on the comparison of the performance of different classification and/or feature selection algorithms. Instead, we choose a method which not only has been shown to reach high

performances in similar speech recognition tasks but whose outcome also is well interpretable linguistically. Random forests (RF) fulfill both requirements. They provide good prediction quality and are able to handle highly correlated feature spaces (Strobl et al., 2008). In the field of speech technology, they are known to perform well in paralinguistic recognition tasks (Schuller et al., 2007), where similarly as in our case, a high number of correlated acoustic features are used as input to the model. Furthermore, RFs have not only been shown to reach high performances with acoustic input, but they also have been shown to outperform traditional n-gram language models (Oparin et al., 2008; Xu and Jelinek, 2004). Given their potential to generalize to unseen data, RF language models are superior to n-grams both in terms of perplexity reduction and WER (Xu and Jelinek, 2004). This is relevant also for our study, as we compare our acoustic-feature based results for homophone disambiguation with the performance of a RF language model, with lexical features as input only.

Only recently conditional interference trees and random forests were applied to linguistic applications (e.g., Tagliamonte and Baayen (2012)). Compared to the very commonly used regression-based approaches in linguistics (e.g., Jaeger (2008)), random forests especially powerful when there are many high-order interactions between the predictors and for problems where the sample size is small but the number of features/predictors is relatively large (Levshina, 2015). Furthermore, predictors can both be categorical or continuous. RFs have been used in several linguistic topics such as modeling speech errors and repairs in spontaneous speech (Plug and Carter, 2014), and turn taking (Roberts et al., 2015). For the statistical analysis of psycholinguistic experiments, Bürki et al. (2011) used random forests to reduce their set of variables before entering them into a mixed effects logistic regression model.

An additional practical benefit of random forests is that they perform both classification and feature selection in an integrated fashion (L., 2001). For feature selection, we do not employ any methods that alter the feature space (e.g., PCA) as we want to investigate the importance of features. Instead, we perform variable selection using tree minimal depth methodology (Ishwaran et al., 2010) as implemented in *randomForestSRC* (Ishwaran and Kogalur, 2015). In this process, we reduce the feature set to 30% of its original size.

3.3. Validation

We evaluate all models using F_1 measures. We use a definition of F_1 which is more suited to multiclass problems:

$$F_1 = \frac{2 \cdot BAC \cdot RR}{BAC + RR}, \tag{5}$$

where BAC is the balanced accuracy or unweighted average, and RR is the overall recognition rate (Batliner et al., 2011). For each word, we perform a speaker-independent tenfold cross-validation and average all folds. We then average the results of all words for our overall results.

A simple way to assess feature importance is to evaluate the *share* and the *portion* of a feature class (Batliner et al., 2011). Share is the number of selected features from a feature class normalized by the total number of features selected. Portion is the number of selected features from a feature class normalized by the total number of features in this feature class.

3.4. Baseline: Lexical model

To evaluate the importance of acoustic features, we compare our acoustic model to a baseline model trained on the target word's lexical context. For this, we train a random forest model on trigrams with the same configurations as for the acoustic model. As previously done for instance in Jurafsky et al. (1998) and Jurafsky et al. (2002), we estimated the log of the conditional probability of a word given the previous words by using a backoff trigram with Good-Turing discounting on the entire spontaneous part of the Kiel Corpus. We decided to train on the entire corpus in order to ensure training material is not limited in size. As the lexical model only uses the left context of the target word, the setting is comparable to real-world ASR where there is no possibility to look-ahead. Furthermore, this model does not receive the POS gold labels of the context as they are not available in ASR either.

4. Results

4.1. Classification performance

Our model trained on acoustic features achieves good overall performance on average (71% F_1, 92% accuracy). In particular, the model trained on lexical features performs considerably worse. While its accuracy of 89% is comparable to the model trained on acoustic features (92% accuracy), its F_1 of 48% is not.

While the reached accuracy is high for all analyzed word forms, this is largely due to strong class imbalance. F_1, however, varies considerably between word forms (see Table 3). The reasons for this variation are diverse. The good performance for ⟨dem⟩ is misleading. Its very strong class imbalance renders any performance highly difficult to interpret.

For other word forms, however, phonetic characteristics are likely to influence performance. Disregarding ⟨dem⟩, the best performance is achieved for ⟨der⟩ (74% F_1). Its diphthong [eːɐ] is very frequently monopthongized in more reduced settings. This is easily detectable when analyzing the spectral features. The worst performance is achieved for ⟨die⟩ (59% F_1). There might be a connection between its performance and ⟨die⟩ being the only open syllable in our analysis (in ⟨der⟩ the coda /r/ is vocalized but not empty). The empty coda is very likely to increase coarticulatory influences from the following syllable (cf. segmental influence discussed in Jurafsky et al. (1998)). Syllable structure and performance for the word forms ⟨das⟩ (61% F_1) and ⟨den⟩ (62% F_1) are comparable.

Table 3: **Results for each word form and for all word forms combined**: Unweighted and weighted average; chance values are given for reference. These are computed by predicting always the most frequent class.

	F_1	*F_1 chance*	ACC	*ACC chance*
⟨der⟩	.740	*.494*	.918	*.954*
⟨die⟩	.596	*.479*	.869	*.851*
⟨das⟩	.610	*.484*	.880	*.880*
⟨dem⟩	.965	*.665*	.993	*.993*
⟨den⟩	.626	*.494*	.954	*.955*
unweighted	.707	*.523*	.923	*.927*
weighted	.664	*.503*	.912	*.916*

Interestingly, accuracies for the acoustic models are approximately equal to chance or in in the case of ⟨der⟩ and ⟨den⟩ slightly worse. We may conclude that the model trained on acoustic features is trading accuracy for F_1: In contrast to accuracy, F_1 is constantly higher than chance (see Table 3).

4.2. Selected features

During feature selection, 57 features were selected. A complete list with all selected features can be found in Appendix A.

4.2.1. Share

The biggest share the selected features is taken by spectral features (28). F_0 features (10), temporal features (9) and intensity features (8) make up the second half of the selected features. Additionally, the word's position in the clause and speaker identity are also included in the final feature set.

4.2.2. Portion

For portion, the situation is inverted, however. Of all feature classes, spectral features achieve the lowest portion. From both temporal features and the *other* feature class (comprised of speaker gender, speaker identity and word position), all but one feature each are selected. From f_0 features and intensity features, about half the features are selected. See Table 4 for details.

Table 4: Share and portion of feature classes.

	# features	# selected	share	portion
spectral	138	28	.491	.203
f_0	19	10	.175	.527
temporal	10	9	.158	.9
intensity	19	8	.140	.421
other	3	2	.035	.667
total	189	57	1	

Notably, the speaker's gender is not included in our final feature set. However, this is unsurprising as speaker identity is one of the selected features. Apparently the differences between homophonic word forms are truly idiosyncratic and are not covered by gender-related generalizations.

In summary, all feature classes employed in this paper contribute to homophone disambiguation. Whereas most selected features are spectral features nearly all temporal features survive feature selection.

5. Discussion

our results reflect the importance of using F1 measure instead of accuracies. hard to compare with results from literature, where acc are shown, but not the chance level nor class imbalances...

Our results indirectly support previous results which showed that acoustic features predict the amount of pronunciation variation in spontaneous speech better than a trigram model Ostendorf et al. (2003). By only analysing the target word's acoustic characteristics we were able to exploit pronunciation variation to classify functions of word forms. Much variation in spontaneous speech is apparently more directly connected to phonetic characteristics than syntax.

Our results also point out some crucial differences between ASR and human speech perception that negatively influence ASR performance. It has been proposed before that there is a link between information on the level of lemmata and phonetically rich information in human speech perception Drager (2011). Our experiments simulate this link and show without strongly diverting from a standard ASR architecture that access to phonetic information benefits the correct identificaiton of lexemes. While humans probable access phonetic detail in earlier stages of speech perception, it is questionable if access to phonetic detail in earlier stages of ASR (e.g., on the level of features or tagged clustering Ostendorf et al. (2003)) would warrant the increase in computational load. Simple rescoring as proposed in this paper allows to avoid this overhead in most cases. It consults phonetic detail only if the target word is easily confusable with another word. Thus, the overhead is rather small in many circumstances. Furthermore, in the current setup, both training and – more importantly – prediction can be done in a reasonable amount of time. This is possible due to the fact that a rescoring system can run in parallel to the core ASR system.

More generally, our results suggest that the complete separation of speech undestanding systems from the acoustic signal is undesirable. A stronger integration of speech recognition and speech understanding may lead to improved output quality for both modules. For this, simple rescoring is likely to be insufficient, however.

5.1. Future work

Our analysis ignored some of the target words' morphosyntactic information – gender, case and number – entirely. All of these, however, are also likely to influence the target words' phonetic realisation. While we do not expect that every combination of word function, gender, case and number is acoustically separable, we believe that some of them are. For instance, from listening and manual inspection we deduce that the differences between

das_{NOM} and das_{AKK} are perhabs negligible. However, it is very much impractical to test any possible comparison due to the resulting very low number of available occurences for less frequent cases.

Research on English function words has shown that not all words are affected identically by changes in environment (e.g., speech rate, segmental and lexical context) Jurafsky et al. (1998). This was also confirmed by our own results as we were not able to perform equally well on all word forms. Therefore, our analysis is limited in as much as it focuses on only five different German word forms. Analysis of more German homophones and homophones in other languages is needed.

6. Conclusions

In this paper, we investigated the possibility of disambiguating German homophones in spontaneous speech using acoustic features. We extracted 7 different feature classes comprising only acoustic features and meta-information (speaker identity, speaker gender, word position) from the Kiel Corpus of Spontaneous Speech, a corpus of naturalistic German dialogues. In total, we extracted 189 features. We showed that in spontaneous German, lexeme-specific information is present in the acoustic signal. We further showed that this information can be used to automatically disambiguate homophones. In our setup, acoustic features generalised well and performed considerably better than a model based on trigrams. Due to the fact that we chose target words with different syllable structure we were able to show that our model's predictive power is dependent on phonological characteristics: We achieve best performance on diphthongues and link this to the fact that the likelihood with which they occur monophthongised is dependent on word function. We achieve worst performance on open syllables and we believe this is due to greater coarticulatory influence from the following syllable which is obviously not correlated with word function.

Our results have important implications for ASR. Phonetic detail can be used in later steps of speech recognition to decrease error rates when homophones are encountered. This is of particular importance for systems dealing with spontaneous speech which contains a high amount of reduction-induced homophones and in languages with high proportion of homophones. Furthermore, in contrast to syntactic analysis, acoustic analysis of homophones is decoupled from general ASR. Thus, it can start before core ASR has finished. Most importantly, our work not only demonstrates novel methods for ASR,

it introduces a new perspective: Whereas previously, the high degree of pronunciation variation in spontaneous speech was primarily seen as a problem for ASR, we view it as an additional resource which is not present in read speech. This change in perspective will guide our future research.

Acknowledgements

The work of Barbara Schuppler was funded by a Hertha-Firnberg grant (T572N23) from the Austrian Science Fund (FWF). The work by Tobias Schrank was funded by the Initial Funding Programme (AF3-442-D1) of the Graz University of Technology.

References

Adda-Decker, M., Lamel, L., 2005. Do speech recognizers prefer female speakers? In: Proceedings of INTERSPEECH. pp. 2205–2208.

Baker, R., Smith, R., Hawkins, S., 2007. Phonetic differences between mis- and dis- in english prefixed and pseudo-prefixed words. In: Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS). pp. 553–556.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., Kessous, L., Amir, N., 2011. Whodunnit – searching for the most important feature types signalling emotion-related user states in speech. Computer Speech and Language 25, 4–28.

Béchet, F., Nasr, A., Spriet, T., de Mori, R., 1999. Large span statistical language models: application to homophone disambiguation for large vocabulary speech recognition in French. In: Proceedings of EUROSPEECH. pp. 1763–1766.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., Gildea, D., 2002. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. Journal of the American Statistical Association 113 (2), 1001–1024.

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., Wellekens, C., 2007. Automatic speech recognition and speech variability: A review. Speech Communication 49, 763–876.

Bickel, B., Comrie, B., Haspelmath, M., 2008. The Leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses.
URL http://www.eva.mpg.de/lingua/resources/glossing-rules.php

Boersma, P., Weenink, D., 2013. Praat: doing phonetics by computer [computer program]. version 5.3.64.

Bürki, A., Alario, F. X., Frauenfelder, U. H., 2011. Lexical representation of phonological variants: Evidence from pseudohomophone effects in different regiolects. Journal of Memory and Language 64 (4), 424–442.

Drager, K. K., 2011. Sociophonetic variation and the lemma. Journal of Phonetics 39, 694–707.

Ernestus, M., Hanique, I., Verboom, E., 2015. The effect of speech situation on the occurrence of reduced word pronunciation variants. Journal of Phonetics 48, 60–75.

Gahl, S., 2008. "time" and "thyme" are not homophones: the effect of lemma frequency on word duration in spontaneous speech. Language 84 (3), 474–496.

Goldwater, S., Jurafsky, D., Manning, C. D., 2008. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase ASR error rates. In: Proceedings of ACL. pp. 380–388.

Goldwater, S., Jurafsky, D., Manning, C. D., 2010. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. Speech Communication 52, 181–200.

Gussenhoven, C., Warner, N. (Eds.), 2002. Papers in laboratory phonology 7. Mouton de Gruyter, Berlin, New York.

Hanulikovà, A., Haustein, J., 2016. Flour or flower? Resolution of lexical ambiguity by emotional prosody in a non-native language. In: Proceedings of Speech Prosody.

Hawkins, S., Smith, R., 2001. Polysp: a polysystemic, phonetically-rich approach to speech understanding. Italian Journal of Linguistics – Rivista di Linguistica 13 (1), 99–189.

Ishwaran, H., Kogalur, U., 2015. Random forests for survival, regression and classification (RF-SRC), R package version 1.6.1.

Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J., Lauer, M. S., 2010. High-dimensional variable selection for survival data. Journal of the American Statistical Association 105 (489), 205–217.

Jaeger, T. F., 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of Memory and Language 59, 434–446.

Jurafsky, D., Bell, A., Fosler-Lussier, E., Girand, C., Raymond, W., 1998. Reduction of English function words in Switchboard. In: Proceedings of the 1998 International Conference on Spoken Language Processing. pp. 3111–3114.

Jurafsky, D., Bell, A., Girand, C., 2002. The role of the lemma in form variation. In: Gussenhoven and Warner (2002), pp. 1–34.

Jurafsky, D., Martin, J., 2009. Speech and language processing. An introduction to natural language processing, computational linguistics, and speech recognition. Pearson Education, Harlow.

Kohler, K., Pätzold, M., Simpson, A., 1995. From scenario to segment. The controlled elicitation, transcription, segmentation and labelling of spontaneous speech. AIPUK 29. IPDS Kiel, Kiel.

Kohler, K. J., Rehor, C., Simpson, A. P. (Eds.), 1996. Sound patterns in spontaneous speech. AIPUK 30. IPDS Kiel, Kiel.

Kohler, K. J., Rodgers, J., 2001. Schwa deletion in German read and spontaneous speech. In: Kohler, K. J. (Ed.), Sound patterns in German read and spontaneous speech: symbolic structures and gestural dynamics. AIPUK 35. IPDS Kiel, Kiel, pp. 97–123.

L., B., 2001. Random forests. Machine Learning 45, 5–32.

Lee, Y.-S., 2003. Task adaptation in stochastic language model for Chinese homophone disambiguation. ACM Transactions on Asian Language Information Processing 2 (1), 49–62.

Levshina, N., 2015. How to do linguistics with R. Data exploration and statistical analysis. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Ligges, U., Krey, S., Mersmann, O., Schnackenberg, S., 2014. tuneR: Analysis of music.
URL http://r-forge.r-project.org/projects/tuner/

Local, J., 2003. Variable domains and variable relevance: Interpreting phonetic exponents. Journal of Phonetics 31 (3-4), 321–339.

Nakamura, M., Iwano, K., Furui, S., 2008. Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. Computer Speech & Language 22 (2), 171–184.

Niebuhr, O., Kohler, K. J., 2011. Perception of phonetic detail in the identification of highly reduced words. Journal of Phonetics 39, 319–329.

Nzgaard, L., Herold, D., Namy, L., 2009. The semantics of prosody: acoustic and perceptual evidence of prosodic correlates to word meaning. Cognitive Science 33 (1), 127–146.

Nzgaard, L., Lunders, E., 2002. Resolution of lexical ambiguity by emotional tone of voice. Memory & Cognition 30 (4), 583–593.

Oparin, I., Glembek, O., Burget, L., Černocký, J., 2008. Morphological random forests for language modeling of inflectional languages. In: IEEE Spoken Language Technology Workshop. pp. 189 – 192.

Ostendorf, M., Shafran, I., Bates, R., 2003. Prosody models for conversational speech recognition. In: Proceedings for 2002 Plenary Meeting and Symposium on Prosody and Speech Processing. pp. 147–154.

Pate, J. K., Goldwater, S., 2013. Unsupervised dependency parsing with acoustic cues. Transactions of the Association for Computational Linguistics 1, 63–74.

Petrov, S., Das, D., McDonald, R., 2012. A universal part-of-speech tagset. In: Proceedings of the International Language and Evaluation Conference (LREC). pp. 2089–2096.

Pierrehumbert, J., 2002. Word-specific phonetics. In: Gussenhoven and Warner (2002), pp. 101–139.

Plug, L., October 2006. Phonetic reduction and pragmatic organisation in Dutch conversation. A usage-based account. Ph.D. thesis, The University of York, Department of Language and Linguistic Science.

Plug, L., Carter, P., 2014. Timing and tempo in spontaneous phonological error repair. Journal of Phonetics 45, 5263.

Rena Nemoto, I. V., Adda-Decker, M., 2008. Speech errors on frequently observed homophones in French: Perceptual evaluation vs automatic classification. In: Proceedings of LREC. pp. 2189–2195.

Roberts, S. G., Torreira, F., Levinson, S. C., 2015. The effects of processing and sequence organization on the timing of turn taking: a corpus study. Frontiers in Psychology 6, 509.

Samlowski, B., Wagner, P., Möbius, B., 2013. Effects of lexical class and lemma frequency on German homographs. In: Proceedings of INTERSPEECH. pp. 597–601.

Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V., 2007. The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In: Proceedings of INTERSPEECH. pp. 2253–2256.

Schuppler, B., van Dommelen, W., Koreman, J., Ernestus, M., 2012. How linguistic and probabilistic properties of a word affect the realization of its final /t/: Studies at the phonemic and sub-phonemic level. Journal of Phonetics 40, 595–607.

Shinozaki, T., Furui, S., 2001. Error analysis using decision trees in spontaneous presentation speech recognition. In: Proceedings of ASRU. pp. 198–201.

Siegler, M. A., Stern, R. M., 1995. On the effects of speech rate in large vocabulary speech recognition systems. In: Proceedings of ICASSP. Vol. 1. pp. 612–615.

Smith, R., Hawkins, S., 2012. Production and perception of speaker-specific phonetic detail at word boundaries. Journal of Phonetics 40 (2), 213–233.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinformatics 9 (307).

Sueur, J., Aubin, T., Simonis, C., 2008. Seewave: a free modular tool for sound analysis and synthesis. Bioacoustics 18, 213–226.
URL http://sueur.jerome.perso.neuf.fr/WebPage$_P$apersPDF/Sueuretal$_B$ioacoustics$_2$00

Tagliamonte, S., Baayen, R., 2012. Models, forests and trees of York English: Was/were variation as a case study for statistical practice. Language Variation and Change 24 (2), 132–178.

Torreira, F., Ernestus, M., 2009. Probabilistic effects on French [t] duration. In: Proceedings of INTERSPEECH. pp. 448–451.

Volín, J., Weingartová, L., Niebuhr, O., 2014. Between recognition and resignation – the prosodic forms and communicative functions of the Czech confirmation tag "jasně". In: Proceedings of Speech Prosody 7. pp. 115–119.

Ward, N., 2004. Pragmatic functions of prosodic features in non-lexical utterances. In: Proceedings of Speech Prosody. pp. 325–328.

Wesener, T., 1999. The phonetics of function words in German spontaneous speech. In: Kohler, K. J. (Ed.), Phrase-level phonetics and phonology of German. AIPUK 34. IPDS Kiel, Kiel, pp. 327–377.

Xu, P., Jelinek, F., 2004. Random forests in language modeling. In: Proceedings of EMNLP04. p. 325 332.

APPENDIX A

Table 5: Set of acoustic features selected for the random forest models. Features are sorted by feature class

feature class	ID	**feature**	**description**

temporal	1	dur	duration of the whole word
	2	s1dur	duration of the first segment
	3	s2dur	duration of the second segment
	4	s3dur	duration of the third segment
	5	s4dur	duration of the fourth segment
	6		
	7		
	8		
	9		
f0	10	frange	
	11	fstdev	
	12	fmad	
	13	fargmin	
	14	fargmax	
	15	fargon	
	16	fargoff	
	17	fslope	
	18	fstderr	
	19	fskew	
	20	fkurt	
intensity	21	irange	
	22	istdev	
	23	imad	
	24	iargmin	
	25	iargmax	
	26	islope	
	27	istderr	
	28	iskew	
	29	ikurt	
spectral	30	ltas1	
	31	ltas2	
	32	ltas3	
	33	ltas4	
	34	ltas5	
	35	ltas6	

	36	ltas23
	37	ltas27
	38	ltas30
	39	ltas31
	40	ltas32
	41	ltas33
	42	ltas34
	43	ltas35
	44	ltas37
	45	ltas38
	46	ltas45
	47	ltas46
	48	plp2
	49	plp3
	50	plp4
	51	plp5
	52	plp6
	53	plp7
	54	plp8
	55	plp9
	56	plp10
	57	plp11
	58	plp12
other	59	position
	60	speaker identity
		token?