Automatic detection of uncertainty in spontaneous German dialogue

Tobias Schrank, Barbara Schuppler

Signal Processing and Speech Communication Laboratory
Graz University of Technology, Graz, Austria
{tobias.schrank,b.schuppler}@tugraz.at

Abstract

Uncertainty is ubiquitous in natural human communication. Human listeners assess the speaker's degree of uncertainty at any time in communication and use this information to shape dialogue. In contrast, currently available computer systems dealing with spoken language are usually not built to perform this task. The ability to detect uncertainty would likely lead to more natural human-computer dialogue. In order to detect uncertainty automatically, we extract linguistic, paralinguistic and dialogue-related features from the Kiel Corpus, a corpus of naturalistic task-oriented spoken German. We then use these features to train a random forests model. Our experimental results show that relatively high classification accuracy can be obtained while employing only 64 well-chosen features (73% accuracy, 69% F_1). To our best knowledge, this is the first study of automatic uncertainty detection using German speech data as well as the first achieving good performance on everyday speech.

Index Terms: uncertainty detection, emotion recognition, conversational speech, spontaneous dialogue, random forests, speech rate

1. Introduction

Uncertainty detection is the task of recognising when humans are uncertain in communication. Information about the speaker's level of uncertainty is present in different channels such as the linguistic content [1], sound [2], facial expressions and gestures [3], posture, etc. Here, we focus exclusively on the analysis of cues that can be extracted from the acoustic channel because in many applications there is only voice input [4]. In speech technology, interest in paralinguistic phenomena such as uncertainty is sparked by the endeavour of making human-computer dialogue more robust and more natural. This would also allow to engage in a greater variety of conversation types [5]. In this regard, uncertainty detection is most important for particular applications: For instance, automatic tutoring systems are required to recognise correct but uncertain answers. The system may thus infer that the student is likely to have guessed the correct answer [4]. Subsequently, when these lucky guesses are treated in the same way as errors this leads to significantly boosted learning performance by the student [6, 7].

Even though in the last two decades automatic uncertainty detection has received growing attention in speech sciences, to our best knowledge the vast majority of research has been limited to a narrow range of conversation types: The speech data is elicited in some form of question-answering task in which one participant is asked factual questions [1, 8, 9, 10, 3, 11, 12, 2, 13, 14, 15, 16, 6, 4, 17, 18]. The question-answering takes place in human-computer dialogue, Wizard-of-Oz dialogue or by reading out answers to written questions. If naturalistic human-human dialogues are used, good performance has not been achieved yet [16]. Moreover, research in uncertainty detection has mainly been studied on English speech data (e.g., all of the above). This may be problematic as paralinguistic phenomena such as uncertainty are well known to take different forms in different languages [19, 20]. In this paper, we attempt to broaden the perspectives in uncertainty detection in two ways. Firstly, we show that good performance on naturalistic human-human conversations that is similar to everyday conversations can be achieved. We manage to do this mainly through careful choice of features. Secondly, we do not analyse English but German speech data. We expect these differences to impact the results of our experiments significantly. In our analysis, we focus on the choice of features and ignore other aspects (e.g., choice of machine learning algorithm). We pay special attention to temporal aspects of the speech signal which are highly informative of uncertainty. In this context, we propose a more detailed model of speech rate. The main aim of the paper is to find an appropriate feature set for the detection of uncertainty in spontaneous German conversations. We then examine this feature set in comparison to feature sets in the literature.

1.1. Feature sets in paralinguistic speech technology

Almost all subfields of paralinguistic speech technology share a common set of core features based on fundamental frequency (f_0), intensity and timing information (e.g., pause duration, utterance duration, speech rate). Typically, the features are descriptors derived from static models of these aspects of the speech signal (e.g., mean, extremes, higher order moments). If at all, dynamic aspects of the speech signal are encoded only indirectly (e.g., position of the extremes, onset, offset) [21]. Frequently, these feature sets are augmented by features common in other speech technology applications [5] or by features for specific tasks (e.g., the nonlinear Teager energy operator in stress detection [22]). Especially in emotion recognition, feature sets are automatically generated [23] which leads to very large feature sets (e.g., the largest feature set in OpenSMILE contains 6552 features [24]).

Previous work has shown that a large variety of features provide information relevant for recognising short-term speaker states [21]. Therefore, we extracted a wide array of features related to the following nine aspects of the speech signal: f_0, intensity, timing, spectral features (MFCCs, formants), voice quality (e.g., jitter, shimmer, HNR, spectral tilt), lexicon (e.g., bag-of-words model, number of hesitations), syntax (e.g., part-of-speech tags, number of clauses and words), dialogue structure (history of previous uncertain dialogue acts) and external features (gender). In contrast to typical feature sets which contain only non-robust descriptors, we derive both robust and non-robust descriptors. Furthermore, speech rate as we model it

differs from other publications. Conventionally, speech rate is modelled as the average number of linguistic units per time unit (e.g., syllables per second) [25]. However, we instead choose to model speech rate as the vector of durations of all syllables in a dialogue act. We then derive features from this vector in the same way as we do for f_0 and intensity (e.g., extremes, mean, standard deviation). We expect the speech rate to be highly informative regarding the speaker's uncertainty. Consequently, we expect the addition of more features related to speech rate to influence the detection of uncertainty positively.

2. Material and annotation

2.1. Kiel Corpus of Spontaneous Speech

The current analysis is based on unprompted, unscripted, spontaneous German speech data from the Kiel Corpus [26]. These data comprises 126 conversations by 18 speaker pairs, each conducting 7 dialogues. They produced a total of 4721 utterances in 2061 turns.

The corpus was collected in a controlled experimental setting with no visual contact between the two test participants. Communication was possible only over an intercom-like device. In order to be heard, speakers had to press a button which muted the interlocutor at the same time. This prevented any overlap of the two participants. The resulting dialogue structure is therefore not entirely representative of human-human dialogue but to some degree similar to human-machine dialogue [26]. In contrast to face-to-face conversations, there are no back-channel phenomena in this speech data.

Speakers were presented with the task of making appointments of different sorts (e.g., a meeting over lunch or a one-week trip). For this, they received made-up time sheets which were prepared to have only a limited number of free time slots in common. This complicated the task of finding suitable time slots and thus speakers were forced to negotiate their appointments. This setup led to fairly naturalistic conversations in which speakers exhibited varying degrees of uncertainty.

2.2. Annotation

The Kiel Corpus contains manual transcriptions and segmentation. For the current study, we additionally segmented each turn into dialogue acts. This was due to the fact that many turns comprise more than one dialogue act. Typically, a turn comprises the second part of an adjacency pair (e.g., the answer to question posed by the interlocutor) and the first part of another adjacency pair (e.g., saying goodbye after completing the task successfully). Aside segmentation, we also annotated each dialogue act by hand on a four-point scale ranging from *certain* to *uncertain*. An example can be seen in Figure 1. These four degrees of uncertainty correspond to possible action performed as response by a dialogue system. These responses can be summarised roughly as: *certain* 0. . . no special action, *rather certain* 1. . . no special action but alerted, *rather uncertain* 2. . . implicit repair, *uncertain* 3. . . explicit repair.

The distribution of uncertainty labels is skewed. Occurrences of labels decrease steadily from *certain* to *uncertain* as the majority of dialogue acts is labelled as *certain* (57%) or *rather certain* (28%). Due to the very low number of *uncertain* instances (0.4%), we merged the instances of *uncertain* and *rather uncertain* to produce a common *uncertain* label (15%).

`UTB [0]:`	Wann würde da bei Ihnen am besten passen? *Which date suits you best?*
`NAR [3]:`	[breathing] Oh, das ist schlecht, weil die Wochenenden bei mir [pause] so ziemlich [pause] ausgebucht sind. Was eigentlich gut ginge, wäre ja, da muß ich jetzt leider mal passen. *Oh, that's problematic because my weekends are pretty much booked up. What would suit me is well unfortunately I have to pass.*
`NAR [1]:`	Wochenendseminar meinen Sie Samstag und Sonntag, ne? *Weekend seminar means Saturday and Sunday, right?*

Table 1: Transcription excerpt with level of uncertainty in square brackets, `NAR`'s turn divided into two parts.

2.3. Validation of uncertainty annotations

A second annotator annotated a small part of the corpus comprising 188 turns (9.1% of the the corpus) following the same annotation manual as the first annotator. For this section of the corpus, the inter-rater agreement between the two annotations lies at 0.539 Kappa which is comparable to or better than other publications in this field [8, 10, 4, 18]. Confusions between the two annotators are mostly involving the label *rather certain* (see Table 2). This may suggest the presence of a mixture of uncertain and certain cues in these dialogue acts. Also, there is a clear bias from the second annotator towards interpreting utterances as more uncertain than the first annotator.

	1st author			
2nd ann.	0	1	2	3
0	162	11	0	0
1	47	59	5	1
2	5	26	34	0
3	0	0	0	0

Table 2: Confusion table for the two annotators.

Our analysis is based only on the second parts of adjacency pairs as annotated by the first author. Our choice to limit the data to the second parts of adjacency pairs stems from two considerations: Firstly, this guarantees better comparability with previous research from the domain of tutoring systems (e.g., [10]). Secondly, this keeps the data homogeneous as the range of pragmatic functions is limited. Thus, for the current study 1158 dialogue acts were selected. This equals to 1813 utterances with a total of 83 minutes.

3. Method

3.1. Feature extraction and normalisation

We extracted acoustic features with Praat [27] and the R packages tuneR [28] and phonTools [29]. Lexical features and features related to dialogue structure are taken from the annotations described in section 2. Part-of-speech tags are estimated with the Stanford POS tagger [30] but mapped to the (simpler) Universal Part-of-speech Tagset [31]. Lexical features are the result of a bag-of-word model. For this, all words are lemmatised and counted. This leads to very sparse feature vectors for lexical features.

All features that constitute counts (e.g., word counts, number of hesitations) are rank normalised by speaker. This has shown to improve performance in terms of both accuracy and computational load [32]. All other features are normalised by speaker using their values' z-scores.

3.2. Feature selection

Especially due to brute-force lexical analysis (1320 features), there is a total of 1430 features. However, many of these 1430 features either provide very little information regarding uncertainty or are highly correlated with other features in the feature set. This issue can be overcome by applying feature-space transformations such as principal component analysis (PCA). However, we do not apply any feature-space transformations in this paper as it essentially prevents any meaningful interpretation of feature importance. We handle the issue of correlated and uninformative features by performing variable selection using tree minimal depth methodology [33]. For this, we use the implementation in randomForestSRC [34]. Following preliminary experiments, we set the absolute number of features to be selected to 64. The final feature set is comprised of 20 timing features, 12 f_0 features, 10 intensity features, 9 spectral features, 5 voice quality features, 4 lexical features, 3 syntactic features and a single feature related to dialogue structure.

3.3. Classification evaluation

We train random forests on the features set resulting from feature selection. For this, we employ the implementation in randomForestSRC [34]. We evaluate the resulting model with a leave-one-speaker-out cross-validation scheme.

In this paper, we present chance-normalised accuracies (CNA) for individual classes in order to detect model bias due to class imbalance [35]. CNA compares a model's performance for each class with the respective relative frequency. It is formally defined as

$$CNA = \frac{p_c - p_e}{1 - p_e}, \tag{1}$$

where p_c is the ratio of correctly classified occurrences and p_e is the expected probability. Note that raw accuracies are always higher than the reported chance-normalised accuracies.

4. Results and discussion

In this paper, we discuss only the relative importance of feature subsets that correspond to the 9 features classes presented in section 3. We choose to not discuss the importance of individual features as the interplay between features is too important to be ignored [21]. The focus on the 9 features classes is theoretically and practically motivated. Theoretically, feature classes hint towards the means humans use to encode uncertainty in speech. Practically, the computational cost of extracting additional features from the same feature class (e.g., minimum and maximum from f_0) is negligible when compared to extracting another feature class (e.g., MFCCs).

4.1. Overall performance

A model based on our feature set of 64 features achieves good overall performance (73% accuracy, 69% F_1). This performance comes about through good performance on the most frequent class *certain* (67% CNA). The performance on the less frequent classes *rather certain* (34% CNA) and *uncertain* (38% CNA) is considerably worse (also see Table 3).

	true		
predicted	0	1	2
level 0	**.916**	.422	.169
level 1	.084	**.490**	.406
level 2	0	.088	**.425**

Table 3: Confusion table for the best model (64 features) showing raw accuracies.

The model's performance is impeded by severe class imbalance. However, preliminary results suggested that neither upsampling nor downsampling increase performance on less frequent classes. Downsampling may be the least preferred option due to the massive reduction of available training data. This alone is likely to decrease performance substantially. Yet, class imbalance is unavoidable in real-life scenarios involving any paralinguistic phenomenon [36].

4.2. Feature classes

A simple way to assess the importance of feature classes is to train a learner on each feature class separately. We evaluate the resulting models the same way as the model built on all features. The model using timing features outperforms all other models significantly (64% F_1). The models using lexical, intensity-related or f_0 features all achieve approximately the same performance ($\sim$55% F_1). The models using one of the remaining feature classes all perform about equally well ($\sim$44% F_1).

Timing features alone perform almost as good (64% F_1) as the best feature set (69% F_1). However, this does not entail that all information regarding uncertainty is encoded in timing features. Only a combination of acoustic, linguistic and external features in the form of the best feature set increases performance (69% F_1). Apparently, all this feature categories carry information about the speaker's degree of uncertainty. For comparison, all acoustic features together only perform as well as timing features alone (both 64% F_1). This suggest a more complex interplay between acoustic and linguistic features.

Remarkably, the class-wise performance differs considerably between acoustic and non-acoustic feature classes. The acoustic feature classes tend to perform particularly well on the least frequent class *uncertain*. Non-acoustic feature classes tend to perform about as well as acoustic feature classes on labels *certain* and *rather certain* (see Table 4).

	acoustic	linguistic	all
level 0	.640	.624	.676
level 1	.112	.049	.143
level 2	.206	$-.091$	.137
features	91	1338	1430
ACC	.674	.649	.692
F_1	.615	.534	.606

Table 4: CNAs for acoustic, linguistic and all features, respectively.

Another way to assess the importance of a feature class is to count the number of members surviving feature selection. In this regard, timing features are clearly the best feature class with 20 out of 64 features. f_0 features (12), intensity features (10) and spectral features (9) perform about equally well. The remaining 11 features contain 5 voice quality features, 4 lexical

publication	corpus	setting	speech type	classes	ACC	F_1	model
Liscombe et al. 2005 [10]	ITSPOKE	tutoring	human-human	3	.76		AdaBoost
Nicholas et al. 2006 [11]	ITSPOKE	tutoring	human-machine	2	.84	.62	AdaBoost
Litman et al. 2009 [13]	ITSPOKE	tutoring	human-machine	2	.66	.60	AdaBoost
Pon-Barry & Shieber 2009 [14]		clozes	read	5	.49		linear regression
		clozes	read	3	.75		linear regression
Dral et al. 2011 [16]	AMI	meetings	human-human	2	.58		decision trees
Forbes-Riley & Litman 2011 [6]	ITSPOKE	tutoring	human-machine	2	.85	.22	logistic regression
this paper	Kiel Corpus	scheduling	human-human	3	.73	.69	random forests
	Kiel Corpus	scheduling	human-human	2	.89	.81	random forests

Table 5: Summary of methods and results as reported in various publications. Reported results are always best results in F_1 if available, and accuracy otherwise.

features, 3 syntactical features and a single feature related to dialogue structure. Thus, 56 features (88%) are acoustic features. This leads to the conclusion that acoustic features are more relevant than linguistic or external features for detecting uncertainty in the current speech data.

4.3. Comparison with previous work

To our best knowledge, this is the first study on spontaneous German that specifically targets the analysis of uncertainty. This allows for a analysis of cross-language differences to previous (English) publications. However, comparison of overall performance of models has to be done with care due to differences in speech data and methodology. In particular, differences in the degree of class imbalance renders the comparison using accuracy almost meaningless. In comparison to those publications that report other metrics than accuracy, the model presented in this paper fares favourably (see Table 5). In the same table, we present raw accuracies to allow for comparisons with all publications listed. Our model with only two classes is added to the table for the same reasons. For this model, we combined the instances of labels *certain* and *rather certain*. This leads to a two-way distinction between *certain* and *uncertain* dialogue acts in the data. Our models outperform all models that present F_1 measures and perform about equally well or better when comparing accuracies.

Previous research put forward different views whether the speaker's gender influences the realisation of uncertainty. Some publications report differences between genders [37, 6] while others negate them [38, 9]. In our analysis, gender is not part of the best performing feature set. This indicates that speaker-dependent normalisation is able to filter out any gender-specific information present in the data.

The performance of some features seems to be tightly connected to the present speech data. For instance, we expected turn-initial silences to be a good indicator for uncertainty. However, due to the use of an intercom speakers apparently tended to produce turn-initial silences in many utterances, not only in uncertain ones. Thus, turn-initial silence is not a good predictor for uncertainty in this speech data and is not part of the best feature set.

Language-specific differences are less pronounced then data-specific differences, however. Leaving obvious differences in the lexicon aside, especially those features performed well that are connected to raised mental effort [39]. These features which are reportedly informative of uncertainty in English are also good indicators for uncertainty in the current German speech data: occurrence of hesitations, timing features and features of duration and length (e.g., utterance duration in seconds and utterance length in words) [4, 40]. As the concept of raised mental effort as put forward in [39] is bound to humans and not to language, these features are only expected to be relatively stable within and across languages.

4.4. Future work

The differences in accuracy between classes (see Table 4) are at least partially due to class imbalance. As neither upsampling nor downsampling improved on this, more elaborate methods are asked for.

By manual inspection of the data, we conclude that information regarding uncertainty is in many cases encoded only locally. Less information seems to be spread over the entire dialogue act. Hence, the analysis of smaller units than dialogue acts appears to be beneficial. A small number of publications in the field of automatic uncertainty detection already did this by analysing words instead of turns [11, 13]. For most applications, the estimates for words have to combined into estimates for the whole turn. We are of the opinion that the methods used for combining labels have not been studied thoroughly enough so far.

5. Conclusions

In this paper, we automatically detected uncertainty in naturalistic spontaneous German human-human conversations. We presented an approach which is based on linguistic, paralinguistic and extralinguistic features. We tested 9 feature classes (timing, fundamental frequency, intensity, spectrum, voice quality, lexicon, syntax, dialogue structure, external features) and evaluated their performance on 1158 dialogue acts taken from the spontaneous part of the Kiel Corpus. The results showed that it is possible to detect uncertainty in speech automatically relatively reliably. The accuracy with which this task is accomplished depended heavily on the feature set employed. In particular, our more complex modelling of speech rate contributed to good classification performance. Automatic feature selection could improve performance even though the machine learning algorithm employed in this paper is built to handle highly correlated features spaces. While only 64 features in size, the resulting feature set outperformed all other feature sets. Even though all features implemented in our system are theoretically motivated and have been used in previous publications, the amount of features that were uninformative regarding the detection of uncertainty in this very speech data is surprisingly large.

6. References

[1] V. L. Smith and H. H. Clark, "On the course of answering questions," *Journal of Memory and Language*, vol. 32, pp. 25–38, 1993.

[2] H. Pon-Barry, "Prosodic manifestations of confidence and uncertainty in spoken language," in *Proceedings of Interspeech*, 2008, pp. 74–77.

[3] M. Swerts and E. Krahmer, "Audiovisual prosody and feeling of knowing," *Journal of Memory and Language*, vol. 53, no. 1, pp. 81–94, 2005.

[4] H. Pon-Barry and S. M. Shieber, "Recognizing uncertainty in speech," *EURASIP Journal of Advances in Signal Processing*, 2011. [Online]. Available: http://asp.eurasipjournals.com/content/2011/1/251753

[5] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Mller, and S. Narayanan, "Paralinguistics in speech and language–State-of-the-art and the challenge," *Computer Speech and Language*, vol. 27, pp. 4–39, 2013.

[6] K. Forbes-Riley and D. Litman, "Benefits and challenges of real-time uncertainty detection and adaption in a spoken dialogue computer tutor," *Speech Communication*, vol. 53, pp. 1115–1136, 2011.

[7] ——, "Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system," *Computer Speech and Language*, vol. 25, no. 1, pp. 105–126, 2011.

[8] ——, "Using bigrams to identify relationships between student certainness states and tutor responses in a spoken dialogue corpus," in *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*, 2005.

[9] E. Krahmer and M. Swerts, "How children and adults produce and perceive uncertainty in audiovisual speech," *Language and Speech*, vol. 48, no. 1, pp. 29–53, 2005.

[10] J. Liscombe, J. Hirschberg, and J. J. Venditti, "Detecting certainness in spoken tutorial dialogues," in *Proceedings of Interspeech*, 2005, pp. 1837–1840.

[11] G. Nicholas, M. Rotaru, and D. Litman, "Exploiting word-level features for emotion prediction," in *Proceedings of IEEE Workshop on Spoken Language Technology*, 2006, pp. 110–113.

[12] K. Forbes-Riley, D. Litman, S. Silliman, and A. Purandare, "Uncertainty corpus: Resource to study user affect in complex spoken dialogue systems," in *Proceedings of the 6th Language Resources and Evaluation Conference (LREC)*, 2008, pp. 513–517.

[13] D. Litman, M. Rotaru, and G. Nicholas, "Classifying turn-level uncertainty using word-level prosody," in *Proceedings of Interspeech*, 2009, pp. 2003–2006.

[14] H. Pon-Barry and S. M. Shieber, "The importance of sub-utterance prosody in predicting level of certainty," in *Proceedings of NAACL HLT: Short Papers*, 2009, pp. 105–108.

[15] ——, "Assessing self-awareness and transparency when classifying a speaker's level of certainty," in *Proceedings of Speech Prosody*, 2010, pp. 100 210:1–4.

[16] J. Dral, D. Heylen, and R. op den Akker, "Detecting uncertainty in spoken dialogues: an exploratory research for the automatic detection of speaker uncertainty by using prosodic markers," in *Affective Computing and Sentiment Analysis: Emotion, Metaphor and Terminology*, ser. Text, Speech and Language Technology 45, K. Ahmad, Ed. Dordrecht: Springer, 2011, pp. 67–77.

[17] H. Pon-Barry and A. R. Nelakurthi, "Challenges for robust prosody-based affect recognition," in *Proceedings of Speech Prosody*, 2014, pp. 144–148.

[18] H. Pon-Barry, S. M. Shieber, and N. Longenbaugh, "Eliciting and annotating uncertainty in spoken language," in *Proceedings of the 9th International Language Resources and Evaluation Conference (LREC)*, 2014, pp. 1978–1983.

[19] G. Ward and J. Hirschberg, "Implicating uncertainty: the pragmatics of fall-rise intonation," *Language*, vol. 61, no. 4, pp. 747–776, 1985.

[20] A. N Chaisaide and C. Gobl, "On the relation between phonatory quality and affect," in *A figure of speech. A festschrift for John Laver*, W. J. Hardcastle and J. Mackenzie Beck, Eds. Mahwah, London: Lawrence Erlbaum Associates, 2005, pp. 323–346.

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

[22] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 201–216, 2001.

[23] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, pp. 1062–1087, 2011.

[24] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proceedings of ACM Multimedia (ACM-MM)*, 2013, pp. 835–838.

[25] I. Lehiste, *Suprasegmentals*. Cambridge, MA: MIT Press, 1970.

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

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

[28] U. Ligges, S. Krey, O. Mersmann, and S. Schnackenberg, *tuneR: Analysis of music*, 2014. [Online]. Available: http://r-forge.r-project.org/projects/tuner/

[29] S. Barreda, *phonTools: Functions for phonetics in R.*, 2014, r package version 0.2-2.0.

[30] K. Toutanova, D. Klein, C. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of HLT-NAACL*, 2003, pp. 252–259.

[31] S. Petrov, D. Das, and R. McDonald, "A universal part-of-speech tagset," in *Proceedings of the International Language and Evaluation Conference (LREC)*, 2012, pp. 2089–2096.

[32] A. Stolcke, S. Kajarekar, and L. Ferrer, "Nonparametric feature normalization for SVM-based speaker verification," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008, pp. 1577–1580.

[33] H. Ishwaran, U. B. Kogalur, E. Z. Gorodeski, A. J. Minn, and M. S. Lauer, "High-dimensional variable selection for survival data," *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 205–217, 2010.

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

[35] B. Wrede and E. Shriberg, "The relationship between dialogue acts and hot spots in meetings," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2003.

[36] A. Batliner, S. Steidl, C. Hacker, and E. Nöth, "Private emotions versus social interaction: a data-driven approach towards analysing emotion in speech," *User Modeling and User-Adapter Interaction*, vol. 18, pp. 175–206, 2008.

[37] M. Lalljee and M. Cook, "Uncertainty in first encounters," *Journal of Personality and Social Psychology*, vol. 26, no. 1, pp. 137–141, 1973.

[38] J. Sherblom and D. D. van Rheenen, "Spoken language indices of uncertainty," *Human Communication Research*, vol. 11, no. 2, pp. 221–230, 1984.

[39] W. J. Levelt, *Speaking: from intention to articulation*. Cambridge, MA: MIT Press, 1989.

[40] N. Ward, "Non-lexical conversational sounds in American English," *Pragmatics & Cognition*, vol. 14, no. 1, pp. 129–182, 2006.