Published in: Langemets, Margit/Penjam, Priit (eds.): The Second Baltic Conference on HUMAN LANGUAGE TECHNOLOGIES. Proceedings. April 4 - 5, 2005, Tallinn, Estonia. - Institute of Cybernetics, Institute of the Estonian Language: Tallinn, 2005. pp. 101-106.

THE "KIEL CORPUS OF READ SPEECH" AS A RESOURCE FOR PROSODY PREDICTION IN SPEECH SYNTHESIS

Caren Brinckmann

Abstract

The naturalness of synthetic speech depends strongly on the prediction of appropriate prosody. For the present study the original annotation of the German speech database "Kiel Corpus of Read Speech" was extended automatically with syntactic features, word frequency, and syllable boundaries. Several classification and regression trees for predicting symbolic prosody features, postlexical phonological processes, duration, and F0 were trained on this database. The perceptual evaluation showed that the overall perceptual quality of the German text-to-speech system MARY can be significantly improved by training all models that contribute to prosody prediction on the same database. Furthermore, it showed that the error introduced by symbolic prosody prediction perceptually equals the error produced by a direct method that does not exploit any symbolic prosody features.

Keywords: text-to-speech, database, CART, symbolic prosody prediction, postlexical processes, duration prediction, F0 prediction, perceptual evaluation, German

1. Introduction

The first text-to-speech (TTS) systems relied mostly on rules that were hand-crafted by human experts. For more than a decade, these hand-crafted rules have been successively replaced by models that are automatically trained on annotated corpora with machine learning (ML) methods. Nowadays, intelligibility is no longer a problem for most TTS systems, whereas naturalness can still be improved. One important factor for natural sounding synthetic speech is the prediction of appropriate prosody, i.e. speech rhythm and melody. In many TTS systems the following modules contribute to the prosodic structure of the generated output:

- *symbolic prosody prediction*: prosodic boundaries, accents (location and type of accent), and intonation contours or boundary tones
- *prediction of postlexical phonological processes*: phonemic deletions, replacements, and insertions (e.g. schwa deletion and assimilation of nasals in German), influencing the rhythmic structure of the synthesised utterance
- *prediction of acoustic parameters*[1]: duration of realised phonemes and pauses, F0 (fundamental frequency) of voiced phonemes.

[1] Other acoustic parameters such as intensity or spectral quality could also be predicted.

Many studies concerning ML-based prosody prediction focussed on the improvement of models for one particular prediction task, e.g. symbolic prosody prediction, duration prediction, or prediction of F0 values (cf. Fackrell et al. 1999). Furthermore, the evaluation of the automatically trained models was mostly corpus-based, comparing the predictions of the respective model with the actual realisations in a speech database. However, the implementation of a specific prediction model into an existing TTS system might not result in *perceptually* improved synthetic speech. For example, Brinckmann and Trouvain (2003) compared ML-based duration prediction (a regression tree) with a rule-based duration prediction model (Klatt rules adapted to German). In terms of "objective" corpus-based evaluation measures (RMSE and correlation), the automatically trained regression tree outperformed the Klatt rules. As long as the input to the duration models was optimal, the regression tree was also perceptually superior to the Klatt rules, but when the models were implemented into the German TTS system MARY (Schröder and Trouvain 2003), the perceptual differences disappeared. The main reasons for this masking effect are the inheritance of error in a complex modular TTS system and the fact that not all models contributing to prosody prediction are based on the same data.

The present study uses the German speech database "Kiel Corpus of Read Speech" (KCoRS) comprehensively for all prosody prediction tasks. The KCoRS comprises over four hours of labelled read speech and is available on CD-ROM (IPDS 1994). As described in Section 2, the original annotation of the KCoRS was extended with additional features that were added mainly with pre-existing tools. On this extended database, several classification and regression trees (CARTs) were automatically trained for all prosody prediction tasks. The corpus-based evaluation measures are given in Section 3. The perceptual evaluation described in Section 4 showed that the output of MARY can be significantly improved by training all models that contribute to prosody prediction on the same database. Furthermore, it showed that the error introduced by symbolic prosody prediction perceptually equals the amount of error produced by a direct method that does not exploit any symbolic prosody features.

2. Database

The textual material of the KCoRS consists mostly of isolated sentences taken from a variety of contexts: train timetable queries, phonetically balanced material, and two very short stories. In total, these are 624 sentences, containing 4932 word tokens and 1673 word types. The recordings of two speakers (male speaker *kko/k61* and female speaker *rtd/k62*), who read the entire material, were used for this study.

The segmental labelling of the KCoRS is essentially phonemic with some phonetic additions (e.g. plosive release phase, glottalisation, and nasalisation). Deviations of the realised form from the lexical phonemes (i.e. deletions, replacements, and insertions) are annotated. Orthography, punctuation marks, as well as sentence and word boundaries are also included in the annotation.

The prosodic annotation incorporates the following domains: lexical stress, accent, intonation contour, prosodic phrase boundaries, and pauses. The accent labels include information about accent location and type (6 categories), degree of accentuation (4 categories), and upstep. The phrase-final intonation contours are labelled with 5 different main categories.

The original annotation was automatically extended with the following features:

- sentence type: statements, exclamations, and 6 different question types
- part-of-speech tags, assigned with the statistical tagger TnT (Brants 2000)
- syntactic phrases of limited depth, assigned with a statistical chunk tagger (Skut and Brants 1998) and the SCHUG parser (Declerck 2002)
- grammatical functions of syntactic phrases, assigned with the SCHUG parser
- word frequency measures from the lexical database CELEX (Baayen et al. 1995)
- syllable boundaries, assigned with a simple algorithm based on standard phonotactic principles of German.

3. Prosody prediction with CARTs

CARTs[2] were trained – using the data of speaker *kko/k61* and *rtd/k62* separately – for the prediction tasks listed in Table 1. For the prediction of postlexical processes and acoustic parameters two types of trees were trained: The first type (*Symbolic*) uses symbolic prosody features, whereas the second type (*Direct*) predicts all segmental features without preceding symbolic prosody prediction (see Figure 1).

Table 1: Mean evaluation measures (across both speakers) for the trained CARTs

prediction task	*Symbolic*	*Direct*
symbolic prosody:	accuracy	
prosodic boundary	95.4%	–
degree of accentuation	88.5%	–
accent location	92.9%	–
accent type	56.2%	–
phrase-final contour	77.8%	–
postlexical processes:	accuracy	
type of change	93.7%	93.1%
replacement rule	93.2%	93.4%
acoustic parameters:	correlation	
duration	0.59	0.56
median F0	0.73	0.64
last F0 in phrase	0.72	0.53

Mean evaluation measures (averaged over both speakers' trees) are given in Table 1. Accuracy of accent type prediction is rather low (56.2%), but closer inspection revealed e.g. that the three peak categories are mostly confounded with one another. The acoustic parameters were predicted as z-scores, so the correlation coefficients are also given in terms of z-scores. For a detailed description of input features, training regime, and results, see Brinckmann (2004).

4. Perceptual evaluation

The corpus-based evaluation measures implicitly assume the realisations of one particular speaker as gold standard. However, usually there are several acceptable ways to produce an utterance, and listeners may have differing idiosyncratic preferences. In order to avoid implementing "improvements" to the TTS system that are not accepted by the listeners, the predictions were evaluated by measuring subjective listener preferences with the Comparison Category Rating (CCR) method of ITU-T recommendation P.800 (ITU-T 1996).

One female and one male diphone-based MBROLA voice (Dutoit et al. 1996) implemented in MARY were used to synthesise 20 sentences. These 20 sentences were randomly selected from the KCoRS and had not been used as training, validation or test items for the CARTs.

[2]All CARTs can be downloaded from http://www.brinckmann.de/KaRS/

Three different prosody prediction methods were used to synthesise each sentence:

- MARY: original MARY system (using hand-crafted rules for prosody prediction)
- *Symbolic*: phoneme identity, duration and F0 values are predicted with CARTs, including intermediate symbolic prosody prediction
- *Direct*: direct prediction of phoneme identity, duration and F0 values with CARTs, *without* using any symbolic prosody features.

In addition, every sentence was copy-synthesised by taking the values for phoneme identity, duration and F0 directly from the respective realisation in the KCoRS.

The TTS architecture used for *Symbolic* and *Direct* is shown in Figure 1. The architecture of the original MARY system is almost identical to *Symbolic* except for some minor differences in the syntactic analysis and the fact that the original MARY system implements all prosody prediction modules with hand-crafted rules instead of CARTs.

The synthesised stimuli were presented to 30 native German speakers by pairs A-B or B-A, where A was copy-synthesised and B used one of the three different prosody prediction methods. The listeners had to judge the overall quality of the second sample relative to the overall quality of the first sample using a 7-point scale (from 3 = *much better* to -3 = *much worse*). The comparison opinion scores (COS), which are presented in terms of the order A-B, were used to compute comparison mean opinion scores (CMOS) for each prosody prediction method, synthesis voice, and listener group.

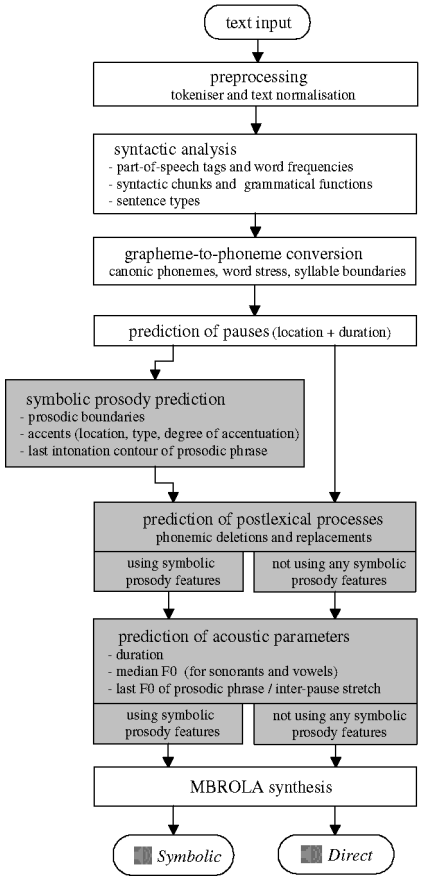

Figure 1: TTS architecture with *Symbolic* and *Direct* prosody prediction. The shaded modules are implemented with CARTs

An analysis of the results (with ANOVA and Tukey HSD) showed that MARY (CMOS -1.55) received significantly lower ratings than both *Symbolic* (-0.76) and *Direct* (-0.80). As shown in Figure 2, only 15.4% of all MARY stimuli have a COS of 0 or higher, whereas 38.9% *Direct* and 39.4% *Symbolic* stimuli are rated having a similar or better quality than the copy-synthesised utterance from the KCoRS. *Symbolic* and *Direct* did not differ significantly for either MBROLA voice.

Listeners having no or little prior experience with speech synthesis generally gave higher ratings (CMOS -0.98) than regular users or synthesis experts (-1.11). CMOS for MARY was especially low for experts/regular users (-1.71).

5. Discussion and outlook

The perceptual evaluation showed that all three prosody prediction methods mostly received negative COS. Thus, one can generally take the copy-synthesised natural utterances as gold standard. However, there are some exceptions where copy-synthesis was rated as inferior to a prediction method. This could be due to idiosyncratic preferences of the listeners or to the limitations of MBROLA synthesis. For example, the two German MBROLA voices do not allow separate modelling of plosive closure and release, even though plosive releases are deleted much more often in German speech than the respective closures. Furthermore, intensity and spectral quality of the concatenated diphones cannot be controlled.

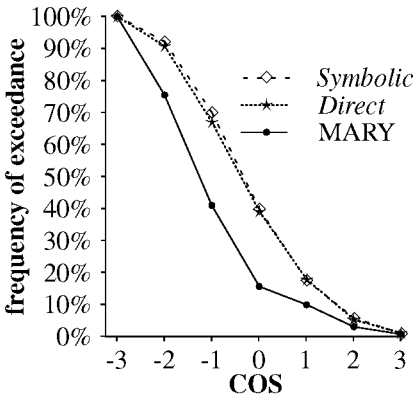

Figure 2: Relative frequency of exceedance of comparison opinion scores (COS), i.e. percentage of ratings that are greater than or equal to the respective COS value, for the three prosody prediction methods

Both ML-based prosody prediction methods *Symbolic* and *Direct* were found to be perceptually superior to the original rule-based MARY method. This shows that the output of a TTS system can be significantly improved by training *all* models that contribute to prosody prediction on the same database.

The two ML-based methods did not differ significantly in the perceptual evaluation. Thus, it can be concluded that the symbolic level of prosody prediction can be safely skipped. On the other hand, the inclusion of symbolic prosody prediction is not detrimental either. The error introduced by symbolic prosody prediction perceptually equals the amount of error produced by the direct method that does not exploit any symbolic prosody features. Therefore, the decision whether or not to include the symbolic prediction can be based entirely on the purpose of the synthesis system. If it is an instructional or research tool (such as MARY), one should include the symbolic prediction level, if it is just a "black box" for the user, one can use the direct prediction method.

As a general rule, the more experienced a TTS user, the higher his or her expectations regarding naturalness. If we aim for a wider usage of speech synthesis, it is necessary to further improve it. More time and effort could be spent introducing other features and trying out different machine learning methods. However, it is doubtful whether the resulting models would lead to a perceptually improved output. The limitations of the KCoRS and MBROLA might have been reached with the presented approach.

One major drawback of the KCoRS is its textual material consisting almost entirely of isolated sentences. In order to model prosodic properties of longer texts, a corpus of read newspaper texts or radio news should be exploited. An even more promising approach is to try a different synthesis method, namely non-uniform unit selection, which generally produces more natural sounding output. However, the available speech material per speaker in the KCoRS is not sufficient for a reliable unit selection speech synthesiser. Therefore, it would be worthwhile to produce such a large labelled speech corpus for German. With this corpus of read speech, one could also include breathing pauses occurring in read speech, making the generated output sound more natural.

References

Baayen, R. Harald; Piepenbrock, Richard; Gulikers, Léon 1995. The CELEX Lexical Database (Release 2). CD-ROM: Linguistic Data Consortium, University of Pennsylvania, Philadelphia, USA [Distributor]

Brants, Thorsten 2000. TnT – a statistical part-of-speech tagger. In: *Proc. ANLP-2000*, Seattle, USA

Brinckmann, Caren 2004. The 'Kiel Corpus of Read Speech' as a resource for speech synthesis. *Master's thesis*: Saarland University: Saarbrücken, Germany. Retrieved February 28, 2005, from http://www.brinckmann.de/KaRS/

Brinckmann, Caren; Trouvain, Jürgen 2003. The role of duration models and symbolic representation for timing in synthetic speech. In: *International Journal of Speech Technology* **6(1)**, 21–31

Declerck, Thierry 2002. A set of tools for integrating linguistic and non-linguistic information. In: *Proc. SAAKM 2002*, Lyon, France

Dutoit, Thierry; Pagel, Vincent; Pierret, Nicolas; Bataille, François; van der Vrecken, Olivier 1996. The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In: *Proc. ICSLP 96*: Vol. 3, Philadelphia, USA. 1393–1396

Fackrell, Justin; Vereecken, Halewijn; Martens, Jean-Pierre; Van Coile, Bert 1999. Multilingual prosody modelling using cascades of regression trees and neural networks. In: *Proc. EUROSPEECH '99*: Vol. 4, Budapest, Hungary. 1835–1838

IPDS 1994. The Kiel Corpus of Read Speech. Volume I. CD-ROM: Universität Kiel, Germany

ITU-T 1996. Methods for subjective determination of transmission quality. ITU-T Recommendation P.800: International Telecommunication Union – Telecommunication Standardization Sector: Geneva, Switzerland

Schröder, Marc; Trouvain, Jürgen 2003. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. In: *International Journal of Speech Technology* **6**, 365–377

Skut, Wojciech; Brants, Thorsten 1998. Chunk tagger – statistical recognition of noun phrases. In: *Proc. ESSLLI-98 Workshop on Automated Acquisition of Syntax and Parsing*, Saarbrücken, Germany

CAREN BRINCKMANN is a research associate at the Institute of Phonetics, Saarland University (Germany). In different projects she contributed to the evaluation and improvement of German speech synthesis systems, multilingual analysis of speech rhythm, and annotation of corpora for investigating information structure. For her master's degrees in computational linguistics and phonetics she focussed on corpus-based prosody prediction for speech synthesis. She offers courses on speech synthesis and programming. Her current research interests include annotation, analysis and modelling of spontaneous speech and corpus-based conversational speech synthesis. E-mail: caren@brinckmann.de.