Conditional Random Fields for Hierarchical Segment Selection in Text-to-Speech Synthesis

Christian Weiss, Wolfgang Hess

Institute for Communication Research
Rheinische Friedrich-Wilhelms University, Bonn, Germany
{cwe, wgh}@ikp.uni-bonn.de

ABSTRACT

In this paper we present the statistically motivated conditional random fields (CRF) approach to concatenative TTS. We use contextual CRFs for speech segment selection where we concatenate the selected segments to an acoustic speech waveform. The CRF approach is used in our corpus-based TTS system AVISS. The acoustic synthesis module consists of trained context dependent CRF models on a multi-level acoustic unit inventory where we apply a hierarchical top-down search to select appropriate segments. The acoustic synthesis is easily adaptable to other languages while there is only the need of a language specific module for text and symbolic preprocessing as well as duration and F0 prediction which can be performed by a prosodic module. The system shows good results in the generated speech waveforms. The CRF approach is usable for acoustic units as well as a parametric synthesis where the speech parameters are generated by CRFs and the speech waveform is produced by a synthesis filter.

Index Terms: Speech Synthesis, Unit Selection, CRF

1. INTRODUCTION

The premise of concatenative TTS is to generate naturally sounding speech by selecting appropriate speech segments from a speech database and concatenating those to an utterance that is predefined by the input text. While there are systems which use sub-word units such as phone-sized or half-phone units [1], others use variable-size speech segments for concatenation. In each case the speech segments have to meet prosodic and spectral requirements or need to be manipulated in order to achieve these requirements. In our system a large speech corpus is used for segment selection where each segment of the speech corpus has its own prosodic and spectral characteristics. The question is how to identify the appropriate potential speech segments for concatenation. To fulfill the requirement of corpus-based TTS with large scale speech data for segment selection a selection strategy is required which enables the system to select the best segment needed in a particular context. There are lots of conditions that must be met in order for such a system to work.

In many successful TTS systems the segment selection algorithm follows a two-dimensional cost function [7] where target costs and join costs are considered. Target cost denotes how close a database unit is to the desired unit. This distance usually follows a Euclidian metric where quantitative, qualitative and prosodic features are included. Join costs are computed according to the question how well two adjacently selected units join together at the concatenation points. This includes a metric in the spectral distance. An overview of spectral measurements can be found in [14]. Both costs are optimized in the sense of finding the best segment in the database which minimizes the overall costs. There is done successful work in trainable TTS where a HMM approach is used for a concatenative TTS. The HMM based systems [3, 6, 13] differ in waveform generation where work by Tokuda et al. (2000) showed that high-quality naturally sounding speech can be produced from speech parameter HMMs itself. This requires training of contextual HMMs using MFCCs as state output vectors and producing the waveforms through a mel log spectrum approximation (MLSA) filter.

In this paper we introduce the CRF approach to text-to-speech synthesis where we use contextual CRFs for speech segment selection. CRFs yield good results on labeling sequential data for instance in part-of-speech tagging [8], as well as in discriminative methods in automatic speech recognition where an improvement using CRFs with hidden states is reported [4]. The process of generating an utterance by concatenating speech segments can also be seen as a sequential production process where the prosodic and spectral features of the speech segments label the given observation and generate a speech segment ID to identify the segment in the speech database. We applied the CRF approach to variable-size segment selection for large scale TTS.

This paper is organized as follows. In Section 2 we review the CRF algorithm and address the use of CRFs in the TTS domain as well as the speech data we used and the features which are considered to train the contextual CRFs. In Section 3 we describe our TTS system where we apply the CRF algorithm to acoustic synthesis. We show the details of the training and the synthesis part. In Section 4 results will be showed and in Section 5 we discuss the described CRF approach and a conclusion and outlook is given on further work.

2. CRF IN TEXT-TO-SPEECH SYNTHESIS

The main application of Conditional Random Fields is in labeling sequential data. Using CRFs for the acoustic synthesis module in TTS will be introduced in this paper. Due to the fact that speech production is a time-series process this can be transformed to a labeling task where speech segments need to receive a label to be identified in the speech database for extraction and concatenation. In the next sections we will review in a short manner the CRF approach and point out the speech database and relevant features used for building contextual CRFs which are utilized to retrieve the appropriate speech segments ID as well as the start and end time of the according segment to generate the output speech.

2.1 Conditional Random Fields

Lafferty, McCallum, and Pereira [8] introduced the conditional random fields approach to labeling sequential data. A conditional random field is a form of undirected graphical model that can be used to define the joint probability distribution over a label sequence given a set of observation sequences to be labeled. Conditional random fields are generalizations of Maximum Entropy Markov Models where the conditional probability of the entire state sequence given the observation sequence is modeled as an exponential distribution. Let *X* and *Y* be jointly distributed random variables respectively ranging over observation sequences *X* to be labeled and their corresponding label sequences *Y*. A conditional random field (*X, Y*) is an undirected graphical model globally conditioned on *X*, the observation sequence. Lafferty et al. define the probability distribution as:

$$P_\lambda(Y \mid X) = \frac{\exp(\lambda \cdot F(Y,X))}{Z_\lambda(X)} \tag{1}$$

$F(y,x)$ represents a set of feature functions $f_1,...,f_n$ and is defined as:

$$F(y,x) = \sum_i \left\langle f_1(y,x,i),...,f_n(y,x,i) \right\rangle \tag{2}$$

where *i* is the index of the speech segment. The features are partitioned due to the stationary assumption and can therefore be weighted according to their importance.

$$Z_\lambda(X) = \sum \exp(\lambda \cdot F(Y,X)) \tag{3}$$

defines a normalization factor where λ is a global weighted vector. The task is now to find the label sequence that maximizes the joint conditional probability which is done by Viterbi search. The parameter estimation of λ is done by a GIS [2] algorithm by maximizing the log-likelihood. There are improvements of the parameter estimation such as described in [15] or [11]. The reader is referred to this work for a detailed view on CRF training.

CRFs are used in various speech processing domains. For instance Sha [11] use CRFs for shallow parsing and Gregory et al. [5] for pitch accent prediction. There are enhancements of CRFs such as dynamic CRFs (DCRF) or Hidden CRF (HCRF) introduced by Gunawardana et al. [4] for phone classification in discriminative methods for automatic speech recognition.

2.2 Speech database

Our TTS system follows a hierarchical segment selection where we select appropriate segments top-down beginning on the word level. While having no segments on the word level the search moves on to the syllable level, the diphone level and finally we select segments on the phone level. This guarantees to minimize the number concatenation points which are always a source for distortion. We are using a 2.5 hour speech database. Table 1 shows its basic properties.

Table 1: Occurrence of data types and tokens in speech data

	Type	Token
Word	6123	40552
Syllable	1497	69037
Diphone	3271	84648
Phone	63	169295

The speech data is recorded at 22 kHz, 16 bit and represents German spontaneous conversational speech which of mainly from the business meeting domain. The originally spontaneous speech data were transcribed and read by a semi-professional female speaker.

2.3 Contextual and dynamic features

For the CRF training we use contextual labels where the feature vector includes the preceding and following token labeled according to their features. We use quantitative features such as positional features, qualitative features such as POS, and prosodic features such as duration and F0. Below we give an overview of the features we used for the contextual labels to train the CRFs.

Table 2: Overview of features

Unit	Feature
Word	preceding, following word sentence type distance left/right in sentence POS duration average F0 mean & std dev of first MFCCs left/right
Syllable	preceding, sFollowing syllable distance left/right word stress duration average F0 mean & std dev of first MFCCs left/right
Phone	preceding, following phone distance left/right word duration average F0

Sentence type as well as part-of-speech tags are accumulated features also used as features on the syllable, diphone and the phone levels.

3. CRF BASED TTS SYSTEM

The CRF based TTS system follows the classical setup where a language specific module serves as a text front end. The text front end normalizes the text and transcribes the given input text into its phonetic representation. The prosodic module predicts duration and F0 and the acoustic module generates the waveform using the CRF models for the respective labels. The system is divided into an offline and online process. During the offline process the context dependent CRFs are trained where all parameters are extracted form the speech database and context dependent labels of the segments are generated. From the given input text the synthesis part generates the waveform by concatenating the segments which are identified by the CRFs.

3.1 Context dependent training of CRFs

During the training part of the system the word, syllable, diphone, and phone segments are labeled. For each segment a context dependent language specific label is generated using the features described in Table 1. Each segment label has as a label class a segment ID which is composed of the utterance ID in the corpus and the according start and end time to extract the segment from the recorded speech data. The CRF training of the acoustic module of the TTS system can be seen as a labeling process like tagging where the label sequence is the segment ID and the feature vector of the segment is the observation. The reader is referred to Section 2. Figure 1 shows a schematic overview of the context dependent CRF training

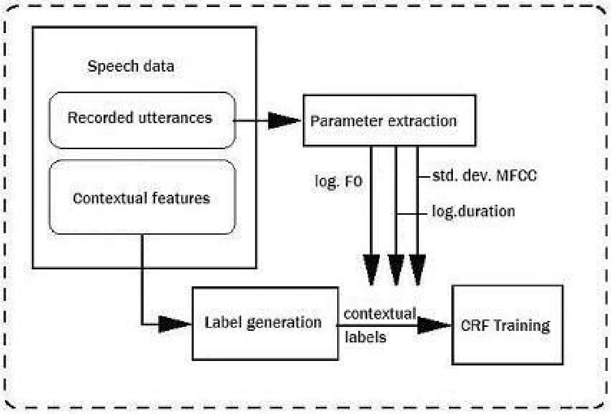

Figure 1: Schematic diagram of context dependent CRF training.

From the annotated speech data we extract the context dependent features as shown in Table 2. Additionally we extract duration and F0 from the recorded utterances. The F0 value is the mean value of F0 over the whole segment. We use logarithmic values of F0 and duration. Further on the MFCC are extracted where we only use the MFCCs from the start and end frame of the respective speech segment and take the mean and standard deviation of the MFCCs as features.

$$F(y,x)=\sum_i \left\langle f_1(y,x,i),...,f_n(y,x,i)\right\rangle \qquad (4)$$

Each feature is represented by a feature function (4).

We include the mean value and the standard deviation of the MFCCs in the training phase. Once the contextual labels are generated for each speech segment the class label is assigned. Here the class label represents the segment ID. For the CRF training we were using the Mallet framework [9]. The training itself is a hierarchical process where the segments are trained independently for each level and for each segment. On the word level there are 6123 context dependent CRFs, one for each word segment. The same procedure is done respectively on the syllable, the diphone, and the phone levels.

3.2 Synthesis by hierarchical segment selection

To synthesize any given input text the system normalizes the text and transcribes the text into its phonetic representation. Sentence types as well as positional features are extracted. A maximum entropy based part-of-speech tagger [10] was trained for German and predicts the according part-of-speech tags for each word. The prosodic module uses a decision tree to predict the logarithmic duration and F0.

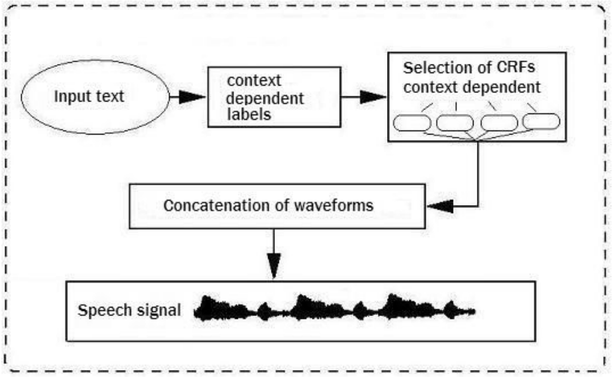
Input text
context dependent labels
Selection of CRFs context dependent
Concatenation of waveforms
Speech signal

Figure 2: Schematic diagram of CRF based segment selection.

Once the contextual label for the word segment has been generated the according CRF is used to provide the segment ID by maximizing the probabilities which is done by a Viterbi search in the according model. Here the cost function approach differs from the CRF where the cost functions minimize the overall costs of the target and join cost function.

If no CRF can be found on the word level we repeat this process on the syllable level. Before using the syllable level the desired token is split in syllables. The syllable boundaries are predicted through a maximum entropy model which was trained for German syllable boundary prediction. Is no according CRF is found on the syllable level next lower level is used until we reach the phone level. Figure 2 shows the synthesis part using context dependent CRFs for speech segment selection.

4. RESULTS

We conducted an evaluation where we compared the speech signals generated with the CRF approach and with conventional two-dimensional cost functions. The listening test was done by students where some had already experience with text-to-speech systems and others not. 34 students took part in the listening test and were advised to rate the speech signals according to their subjective impression. The categories they should rate are:

- General impression
- Naturalness
- Intonation
- Distribution of pauses and phrase boundaries
- Quality (absence of distortions)

We used a 1-5 scale rating system where 1 is worst and 5 are best. Figure 3 shows the mean values of the categories. The white bar represents the conventional approach and the black bar the CRF based approach. The entire figure shows that in each category, except pauses, the CRF approach had better evaluation results than the conventional one.

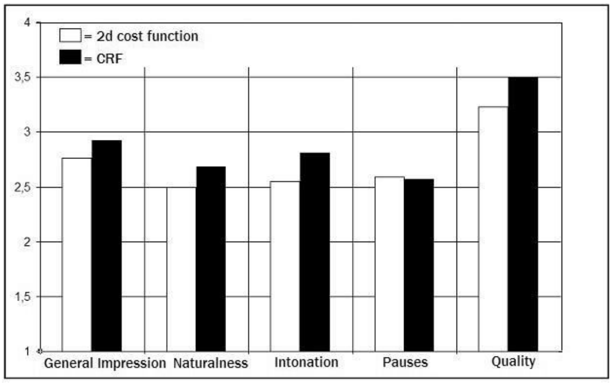

Figure 3: Overview of listening test

5. DISCUSSION AND CONCLUSION

Unit selection based text-to-speech synthesis from large speech databases produce high quality speech. There are some disadvantages using this kind of synthesis technique. One is the effort to setup a database. The annotation of the database could most of the time done with tools which automate the process, but there is always a manual correction needed if the speech database should be of high quality. The second disadvantage is the resource. Large speech databases need an amount of storage memory. It makes this kind of TTS unusable on mobile devices. On the other hand those kinds of TTS systems can produce highly natural speech.

By now most of the systems use the two dimensional cost function approach. There is some work done using statistically motivated approaches for speech segment selection. We described our CRF based approach to speech segment selection. The CRF approach was integrated in the acoustic module of our software. After doing some tests it showed that the CRF approach gave better results than the conventional cost function approach. Speech samples can be downloaded from our webpage http://www.ikp.uni-bonn.de/~cwe. The training of the contextual CRFs took about one week on a 2.4 GHz Pentium 4 PC. This training could be speeded up using efficient training algorithms. During runtime the selection process is slower than our standard unit-selection system.

The CRF approach is a promising approach to speech segment selection. It selects that speech segment which maximizes the conditional joint probability and is therefore more often precise than the cost function approach where it is not clear how the features should be weighted to achieve the best selection strategy. The CRF based approach can also be used to train context dependent CRFs with speech parameters as class label and to use a synthesis filter to generate the speech waveform. Future work will investigate the use of CRFs to generate speech from speech parameters as well as steps towards improving the speed of the CRF based acoustic synthesis module.

6. REFERENCES

[1] Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y. and Syrdal, A., "The AT&T Next-Gen TTS system", *Joint Meeting of ASA, EAA, and DAGA*, 1998.

[2] Darroch, J. N. and Ratcliff, D., "Generalized iterative scaling for log-linear models," *Ann. Math. Stat., vol. 43, no. 5*, 1972.

[3] Donovan, R. E., "Trainable Speech Synthesis", *PhD Thesis*, Cambridge University Engineering Department, 1996.

[4] Gunawardana, A., Mahajan, M., Acero, A, and Platt J. C. "Hidden conditional random fields for phone classification", in *International Conference on Speech Communication and Technology,* 2005.

[5] Gregory, M., Altun, Y., "Using CRF to Predict Pitch Accent in Conversatonal Speech", in Proc. *42nd Annual Meeting of the Association for Computational Linguistics* (ACL), 2004.

[6] Huang, X., Acero, Hon, H., Ju, Y., Liu, J., Meredith, S. and M. Plumpe, "Recent improvements on Microsoft's trainable text-to speech system - Whistler," *Proceedings of ICASSP*, 1997.

[7] Hunt, A. J. and Black, A. W., "Unit selection in a concatenative speech synthesis system using a large speech database", *Proceedings of ICASSP*, 1996.

[8] Lafferty, J., McCallum, A. and Pereira, F. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", *Proceedings of ICML*, 2001.

[9] McCallum, A. K., "MALLET: A Machine Learning for Language Toolkit", http://mallet.cs.umass.edu , 2002

[10] Ratnarparkhi, A.: "Maximum Entropy Models for Natural Language Ambiguity Resolution". *PhD Dissertation*, University of Pennsylvania: 1998.

[11] Sha, F., Pereira, F. "Shallow parsing with conditional random fields", Proceedings of HLT-NAACL, 2003.

[12] Stöber, K., Wagner, P., Klabbers, E., Hess, W., "Definition of a training set for unit selection-based speech synthesis", Proceedings of Speech Synthesis Workshop 4 , Scotland 2001.

[13] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi T. and Kitamura, T., "Speech parameter generation algorithms for HMM based speech synthesis", *Proceedings of ICASSP*, 2000

[14] Vepa, J., King, S. and Taylor, P., "Objective distance measures for spectral discontinuities in concatenative speech synthesis," *Proceedings of ICSLP*, 2002.

[15] Wallach, H., "Efficient training for Conditional Random Fields" *Master Thesis*, University of Edinburgh, 2002.