The DRESS Blizzard Challenge 2008 Entry

Hamurabi Gamboa Rosales, Oliver Jokisch and Ruediger Hoffmann

Laboratory of Acoustics and Speech Communication, Dresden University of Technology, Germany
[Hamurabi.Gamboa,Oliver.Jokisch,Ruediger.Hoffmann]@ias.et.tu-dresden.de

Abstract

This paper presents the version of the DRESS System that participated in the Blizzard Challenge 2008. Here we show a basic overview of the DRESS system, methodology and a description of its unit selection. We present a novel concept to increase the efficiency of the exhaustive speech unit search within the database by an unit selection model that is based on mapping analysis of the concatenation costs and Bayes optimal classification (BOC).

Index Terms: speech synthesis, unit selection, join costs

1. Introduction

Recent improvements of the Dresden TTS system (DRESS) for participating in Blizzard Challenge 2008 are described in the present paper. DRESS is a corpus-based time-domain synthesizer with pre-processing module, grapheme-phoneme converter, duration control, intonation control, unit selection module and acoustic module. It is available as a software system. Recent improvements refer to multilingualism and naturalness. Multilingualism is obtained by a dedicated structure that can handle databases from different languages. Databases for German, English, Spanisch, Czech and Chinese have been developed. Naturalness is a highly important feature of synthetic speech. Apart from the segmental quality and the voice characteristics, it depends mostly on the prosody. For this purpose, DRESS was equipped with an intonation module (neural-network based approach). Additionally, a novel unit selection module is applied to avoid mismatches and distortions, which degrade significantly the quality of the synthesized speech signal. Mismatches are known as concatenation cost, which could be considered as an estimator of the quality of speech synthesis. If the discordance between a speech unit and the predicted specification is also taken into account, the quality of the synthesized speech signal even suffers an extra degradation (target cost). Therefore, it is necessary to set up all these factors in one integrated function, which represents the influence of target and concatenation costs on the resulting speech synthesis quality and enables the finding of optimal speech units sequences to obtain the desired synthesized waveform. But the processing of all information requires an exhaustive training to set up the weighted coefficients for both sub-costs [1][2]. Therefore, we present an unit selection framework based on Bayes optimal classification (BOC) and its experimental evaluation. BOC has a principle advantage because it does not require an exhaustive training to set up weighted coefficients for target and concatenation sub-costs. It can provide an alternative for unit selection but requires further optimization, e. g. by integrating target cost mapping.

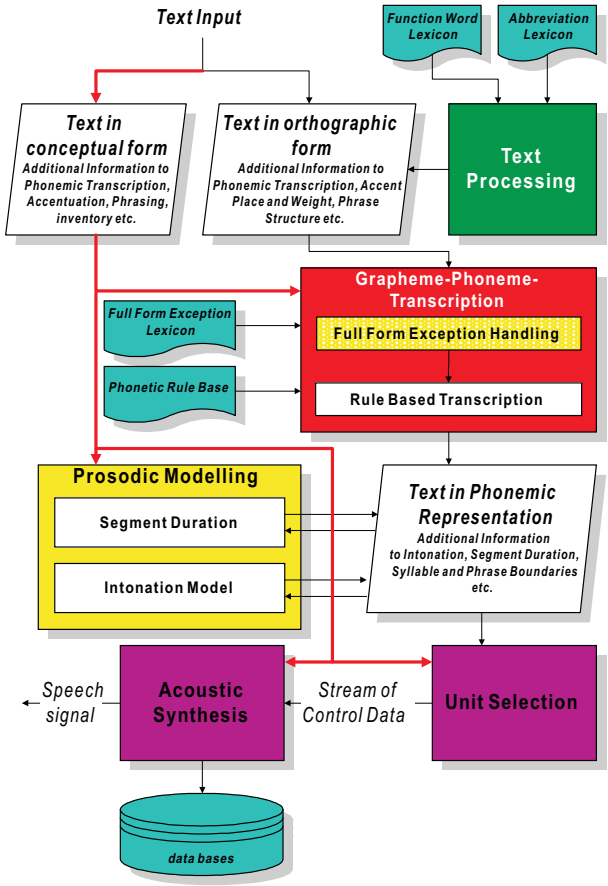

Figure 1: *Schematic diagram of DRESS.*

2. The TTS SYSTEM DRESS

This chapter gives an overview of the **Dres**den **S**peech **S**ynthesis System (DRESS) as shown in Figure 1:

2.1. Text processing

Fig. 1 presents the components for processing at symbolic level. At first, the input text is split into sections. Plain ASCII-text or text enriched with conceptual information - containing pronunciation forms of some words, accent- or boundary-tags - is processed by the pre-processor stage. Word, phrase and sentence boundaries are classified and tagged, special character combinations are recognized, function words and abbreviations are detected and marked in the running text. The sentence boundary detection determines the type of all sentences. The rules for processing numerical formats and text anomalies like abbrevia-

tions and names are tried first, to generate a phonemic form of the input text. On all remaining text parts, a tokenizer is applied. By means of a special lexicon, function words are processed. If the process was not yet successful, the abbreviation lexicon is applied to the tokens. The phonemic rule set and afterwards the spelling module are applied to convert the input text. Finally, phrase boundary detection yields a finer subdivision of the sentences into prosodic phrases. The Grapheme-to-Phoneme stage derives a phonemic representation of the input text. Processing is done first by a lexicon-based and then by a rule based component. The construction of the phonetic rule base follows an approach described in [3]. All phonetic rules are organized in graphemic prefix, rule body, suffix and a phonemic result. Furthermore, accent type and place are supplied to the following prosodic components.

2.2. Prosodic modelling

Prosodic processing (duration and intonation control) is done to the stream of phonemic information by several modules like add segmental durations and pitch parameters .

2.2.1. Intonation model

Data Driven (Neural) Approach is utilized to generate flexible speaking styles and to quickly adapt the DRESS system to the requirements of different voices or languages - a data driven approach is used. This approach includes a artificial neural network (ANN) and enables the direct estimation of the f0 contour from a sequence of linguistic input vectors. The feature coding is syllable-oriented: From the phoneme sequence, syllables will be isolated and stepwise presented to a recurrent network. For each syllable, a vector of N1 = 8 linguistic and phonetic features is applied to the network input. The first hidden layer consists of N2 = 10, the second hidden layer of N3 = 6 neurons. The second hidden layer is completely connected to the context neurons, i.e. the ANN input layer contains C*N1+N3=46 neurons. The output layer owns N4 = 3 neurons, which estimate the f0 contour of the focus syllable. The input encoder considers the phrase position, stress situation, phonetic features of the nucleus and its context (see Fig. 2).

2.3. Unit Selection

The unit selection attempts to find the best combination of unit sequences to assure that the perceptual differences between expected (natural) and synthesized speech signal are as low as possible. It transforms the stream of phonemes into a sequence of speech units and joins it with the prosodic information. We propose a novel unit selection algorithm, which will be discussed in more detail in the following chapters.

2.4. Acoustic synthesis

Finally, the acoustic synthesis builds up a synthetic speech signal from the sequence of speech units and reproduces the prosodic parameter contours. Additionally, it may apply some prosody manipulations to fundamental frequency (f0) contour and duration of phonemes to the speech signal. Basically, both mechanisms behave in a contrary manner. The duration of a phoneme is the sum of the duration of its periods, which are obviously changing while the fundamental frequency is manipulated. This implies that f0 cannot be modified independently of the duration. That is why manipulating the prosody of a speech signals should be an iterative process. Both frequency and duration have to be shifted step-by-step to their target values [4].

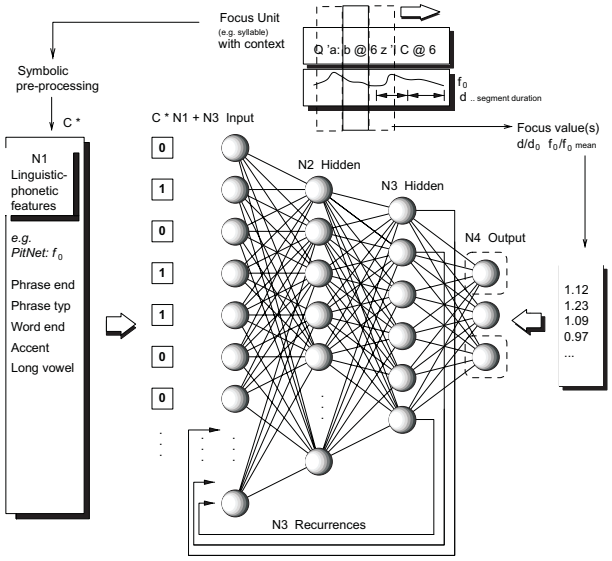

Figure 2: *ANN model for intonation control DRESS.*

3. Unit Selection Framework

The target and concatenation costs have been integrated in a total cost function by [1], which represents the degradation on a synthesized speech signal. Additionally, they described an unit selection model as a search for a low cost candidate unit sequence. Although, different target and concatenation sub-costs have been proposed to unit selection, the feature sub-costs like duration, f0, energy, linear spectral frequencies (LSFs), multiple centroid analysis (MCAs) [5] and Mel frequency cepstral coefficients (MFCCs) have already reached a significantly representation of the deterioration of a synthesized signal. Hence, they compose a special unit selection process in such a way that the sum of the target and concatenation costs determines the total cost C for a sequence of n speech units.

$$C(t^n, u^n) = \sum_{i=1}^{n} \sum_{j=1}^{p} w_j^t c_j^t(t_i, u_i) \qquad (1)$$

$$+ \sum_{i=2}^{n} \sum_{j=1}^{q} w_j^c c_j^c(u_{i-1}, u_i)$$

Where t_i represents the searched predicted specification, u_i the speech unit, c_j^t target cost, p the number of weighted target sub-costs, c_j^c concatenation cost, q the number of concatenation sub-costs and w the weighted coefficients (WCF). The following step should be to find those weighted coefficients that determine the effect-weight of every target and concatenation sub-cost in the total cost function. This is considered as the best way to find the right speech unit sequence for the desired synthesized speech signal. However, the search for the optimal weighted coefficients is not a trivial task, because it normally requires training, which is a subjective work and time consuming for every speech database [1][6][7][8]. Therefore, we present a unit selection framework that is based on mapping of the concatenation sub-costs and a Bayes Classifier. Therewith we avoid principally the exhaustive and subjective search of weighted coefficients. Also, we estimate in great part the quality or degradation of the synthesized signal by mapping the concatenation sub-costs.

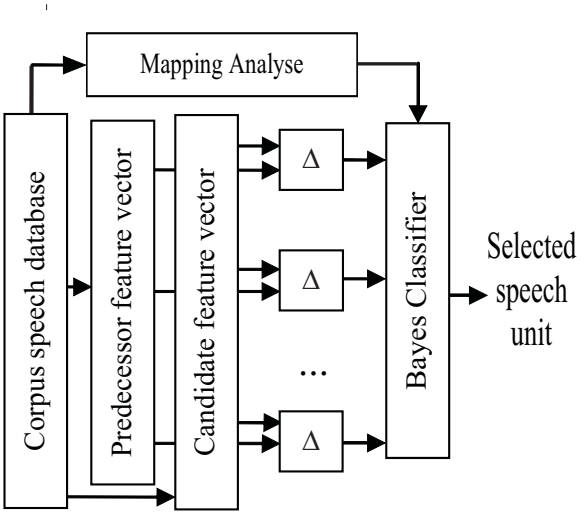

Figure 3: *Proposed Unit Selection Framework.*

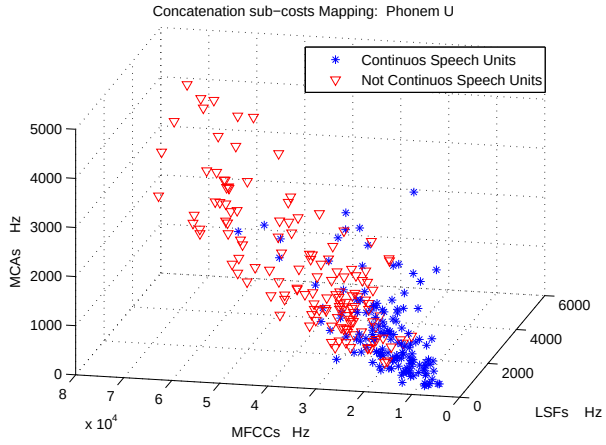

Figure 4: *Concatenation sub-costs mapping.*

3.1. Bayes classification framework

The Bayes classification framework is composed by different modules that are shown in Fig. 3. It illustrates the speech database, where all possible speech units that compose the desired synthesized speech signal are searched for. The speech unit candidates are chosen in the speech database by Backward Oracle Matching algorithm (BOM) [9]. It picks up all possible speech units that compose the phonetic sequence of the text to be synthesized. Once the speech units are found, their MCAs, LSFs and MFCCs coefficients are calculated at the right and left boundaries and represented in a vector form. Afterwards the distance Δ between predecessor and candidate speech unit sequence of the desired synthesized speech signal is calculated. The mapping is obtained by calculating the concatenation sub-costs distance of the speech units. Finally, the Bayes classification determines if the concatenation between the speech units is corrupt or proficient. In the following sections the components of the proposed unit selection framework will be described in more detail.

3.1.1. Parametric distance function

The Delta symbols in Fig. 3 show the distance function. They compare two values of the same feature and produce a distance value output. This function measures the degree of match between the features of two adjacent speech unit candidates. The distance is calculated with 20 ms frames, 9 MCAs, 26 LSFs and 24 MFCCs coefficients features vector in the corresponding boundary at the point of concatenation. We utilize the Mahalanobis distance measure, because it has shown a high correlation with human perception of discontinuity at concatenation boundaries [10].

$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T K^{-1}(\vec{x}-\vec{y})} \qquad (2)$$

Where $\vec{x}$ and $\vec{y}$ are the features vectors of the predecessor and candidate speech units and $K^{-1}$is the inverse covariance matrix. That is how the distance between the speech units for the MCA, LSF and MFCC features is calculated .

3.2. Mapping Analysis

The mapping consists of an off-line calculation of the concatenation sub-costs between speech units in the database, which do and do not present displeasing distortions when they are concatenated. It is estimated by the distance calculation between the speech unit features like MCAs, LSFs and MFCCs at the right and left boundaries. The mapping of concatenation sub-costs that do not present any distortion is done by the concatenation sub-cost distance calculation between speech units, which are continuous in the words or sentences in the speech database. Although, the concatenation sub-costs of continuous speech units are set up to zero by definition [1][2], we utilize the calculated concatenation sub-cost distances to map the real values of continuous speech units off-line. In this way, we obtain a real reference of concatenation sub-costs without distortion. The mapping of the concatenation sub-costs that present unpleasant distortions is done by using a determined set of speech units. These speech units come from different words or sentences contained in the speech database and were previously selected to not concatenate properly. Therewith the second reference is also obtained. We were able to differentiate between two mapped references, which represent the proficient and corrupt areas of concatenation as it is shown in the Fig. 4. It illustrates the mapping of the concatenation sub-cost distances between continuous and not continuous speech units at the point of concatenation. For this instance the concatenation type at the middle of a short vowel /U/ is shown, because the concatenation between short vowels has proven to be the most inclined case to concatenate not properly [11][12]. Finally, a mapping for every phoneme concatenation should be done. Consequently, the next task is to determine the concatenation sub-cost area, which can determine if the join between two not continuous speech units is a proficient or corrupt concatenation based on the corresponding previously mapping pro phoneme by a classification method like BOC.

4. Bayes Optimal Classification

Bayes optimal classification establishes that the class probability k given the feature vector $\vec{x}$ is equal to multiplication between the a priori likelihood the class $P(k)$ and the density probability function $P(\vec{x}/k)$ divided by the probability of the sample, according to equation (3).

$$P(k/\vec{x}) = \frac{P(\vec{x}/k) \cdot P(k)}{P(\vec{x})} \tag{3}$$

Where k is the proficient or corrupt concatenation class and $\vec{x}$ is the concatenation sub-cost distance vector between two speech units. The denominator is not considered, because it is common to both concatenation classes. A priori probabilities of continuous and not continuous concatenation sub-costs have been assumed equal 0.5. Also, we assumed the independence between feature vectors, so that the BOC combines the impact and probability of feature vector on the class label. The BOC was modeled with a multivariate density Gaussian distribution [13] considering that the feature vectors have a normal distribution as is showed in the following equation (4).

$$P(\vec{x}/k) = \frac{1}{(2\pi)^{N/2}\,|K_k|^{1/2}} \cdot \exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu_k})^T K_k^{-1}(\vec{x}-\vec{\mu_k})\right] \tag{4}$$

Where $\vec{x}$ is the Mahalanobis distance by concatenating speech units, the covariance matrix K and mean μ are calculated according to the class feature vectors of the Mahalanobis distance. Afterwards we would like to find those speech units that have the maximum probability. It is achieved by a discriminant function as it is described in the following equation (5) and (6) .

$$e = \underset{i=1,\ldots K}{\arg\ max} \qquad d_i(\vec{x}) \tag{5}$$

$$d_i(\vec{x}) = P(k) \cdot P(\vec{x}/k) \tag{6}$$

Where e is the maximum argument of the discriminant function $d_i(\vec{x})$ in the equation (6), which contains the maximum probability. K is the number of classes (proficient or corrupt concatenation).

4.1. Bayes discriminant function

By the substitution of the multivariate density Gaussian distribution (4) the discriminant function (5), we obtain the corresponding distance Bayes discriminant function (7) as it is shown in Fig. 5. The discriminant function allows to classify a concatenation between two not continuous speech units into corrupt and proficient concatenation type, which is based on its probability estimation.

$$d_k^*(\vec{x}) = \ln\left[d_k(\vec{x})(2\pi)^{N/2}\right] \tag{7}$$

$$d_k(\vec{x}) = \ln P(k) - \frac{1}{2}\ln|K_k| - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} K_k^{(mn)}\mu_{mk}\mu_{nk} + \sum_{n=1}^{N}\Big(\sum_{m=1}^{N} K_k^{(mn)}\mu_{mk}\Big)\cdot x_n - \frac{1}{2}\sum_{m=1}^{N} K_k^{(mn)}x_m^2 - \sum_{m=1}^{N-1}\sum_{m>m}^{N} K_k^{(mn)}x_m x_n$$

Equation (7) describes a Bayes discriminant function [13], which can be used to calculate the corresponding discriminant function for every concatenation of phonemes of not continuous speech units in the speech database. The Bayes discriminant function at the point of concatenation of the nasal phoneme /N/ is shown in the Fig. 5. It illustrates 2-Dimensional mapping analysis, where the both concatenation areas are delimited by

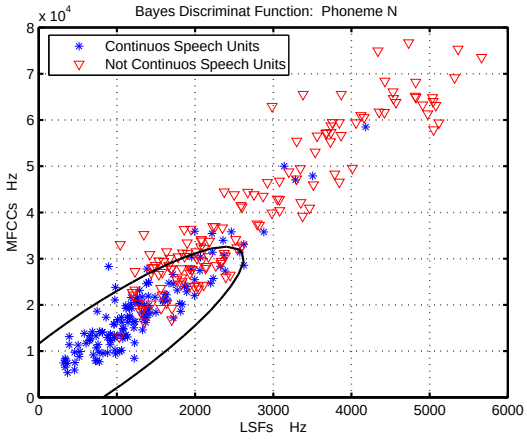

Figure 5: *Bayes discriminant function.*

Bayes discriminant function. It is easy to recognize that some concatenation sub-costs of not continuous speech units fall inside the proficient concatenation area, which is known as classification error [13].

4.2. Unit selection process

Firstly, the speech units that had been found for the desired synthesized speech signal by the BOM are processed by the Bayes Classifier. The BOC classifies the speech units in corrupt and proficient concatenation types by using the corresponding discriminant function, as it is shown in Fig. 5. Then, the speech units, which were found to concatenate corrupt and whose concatenation sub-costs do not fall into the proficient concatenation delimited area by the discriminant function, are removed from the unit selection process. Afterwards, the left over speech units are computed by the corresponding previously obtained distribution of the proficient concatenation type by using maximum likelihood method. Finally, the concatenation of speech units that shows the highest likelihood is selected. In this way, these speech units are selected that match best with the searched phoneme sequence to obtain the desired speech signal without distortions by the concatenative-based speech synthesis.

5. Listening Test

The DRESS TTS system [14] was used to synthesize three blocks of 10 utterances with three different unit selection methods. Furthermore, the previously mentioned speech database "TC-STAR" was utilized for the three unit selection methods in the DRESS TTS system. The first block ("Conventional US") was synthesized by the unit selection method proposed in [1], which represents the basic principles of the sum of target and concatenation costs for unit selection and requires an exhaustive training to set up the weighted coefficients for target and concatenation sub-costs. By the second block ("Masking US") the unit selection proposed in [15] was used, which bases its unit selection method on previously defined transparency and quality functions and determines if a determined concatenation will or will not present distortions. The last block ("BOC US") is our proposed unit selection method.

5.1. Experiment

The synthesized utterances were evaluated by 10 listeners and finally a mean opinion score (MOS) of their absolute category decisions has been calculated. All listeners were students or researchers at Dresden University of Technology with good English proficiency, experience on speech recognition and synthesis. Their age varied from 20 to 30 years. The listening test consisted of the evaluation of intelligibility, naturalness and concatenation quality of the synthesized utterances. The probands listened to the test stimuli in random order. We asked them to rate the quality of the synthesized utterances on a scale of 1 (Bad) to 5 (Excellent). The MOS values obtained for the three unit selection methods are summarized in Table 1.

Table 1: *Mean opinion scores.*

Conventional US	Masking US	BOC US
2.76	2.25	2.67

5.2. Results

The mean opinion scores in Table 1 turned out to be significant at the one percent-level by paired t-test. The masking method of unit selection based on the masking quality function has obtained the worst results in the listening test. This is due to the quality concatenation masking function that can not be determined by a linear function for every type of concatenation as it was proposed by [15]. The conventional unit selection method, which is based on the sum of target and concatenations costs, performed slightly better than BOC. It reflects the potential improvements that can be obtained by taken into account the target sub-costs in the speech synthesis. Nevertheless, the task of setting up the weight coefficients on the total cost C function in the equation (1) was a very difficult subjective work, which required many hours of listening training for the specific corpus database. Summarized, the proposed BOC unit selection obtained better results than the proposed masking method of unit selection and it was slightly worse than the conventional unit selection method manifesting only a small perceptive difference between them. BOC unit selection performance is functional since it has shown an acceptable quality and avoided many hours of training to determine an appropriate search for the best speech unit sequence by mapping the concatenation sub-costs, which is mainly considered as a subjective task.

6. Conclusion

The participation at Blizzard Challenge 2008 was a great benefit to the DRESS system. It gave us the opportunity to present another perspective on unit selection methods for corpus-based speech synthesis by proposing a Bayes optimal classifier. BOC unit selection is based on concatenation and sub-costs mapping of speech units representing distortions in the concatenated unit sequence. In this method, the mapping provides two references of proficient and corrupt concatenation areas. Furthermore, a discriminant function as shown in the equation (7) was developed, which calculates the probability estimation of proficient and corrupt concatenation type between two speech units by this discriminant function. BOC unit selection provides a good unit selection alternative with a similar or better performance than the available unit selection methods of this TTS system. BOC has one principle advantage because it does not require an exhaustive training to set up the weighted coefficients for target and concatenation sub-costs. Therefore, BOC unit selection supports the integration of new speech databases like the database provided by Blizzard Challenge 2008 in a TTS system avoiding exhaustive training for each newly integrated speech database. In future, it will be important to improve the BOC unit selection performance by the integration of target cost mapping and to participate in the following Blizzard Challenge, because the target cost has a great influence on the naturalness of the synthesized speech signal.

7. References

[1] Hunt, A.J. and Black, A.W., “Unit selection in a concatenative speech synthesis using a large speech database”, in Proc. ICASSP, pp. 373-376, 1996.

[2] Beutnagel, M., Conkie, A. and Syrdal, A.K., “Diphone synthesis using unit selection”, Proc. 3rd ESCA/COCOSDA International Workshop on Speech Synthesis, Jenolan Caves, pp. 185-190, 1998.

[3] Wothke, K., MA., “Letter-to-phone rules for German”, Technical Report 75.91.04, Feb. 1991, IBM Heidelberg Scientific Center.

[4] Hoffmann R., Jokisch, O., Hirschfeld, D., Strecha, G., Kruschke, H., Kordon, U., “A multilingual TTS system with less than 1 megabyte footprint for embedded applications”, In Proc. Int. Conference on Acoustics, Speech and Signal Processing (ICASSP), HongKong, China, April 6-10, 2003, vol. 1, 532 - 535.

[5] Corwe, A. and Jack, MA., “Globally optimizing formant tracker using generalized centroids”, Electronic Letters, Vol 23, No. 19, pp 1019-1020 Beijing,China, 1987.

[6] Gamboa Rosales, H. “Evaluation of smoothing methods for segment concatenation based speech synthesis”, In Proc 16th Czech-German Workshop “Speech Processing”, September 77-83, Prague, Czech Republic, pp. 270-273, 2006.

[7] Alas, F., Llor, X., Formiga, L., Sastry, K., Goldberg, D. E., “Efficient Interactive Weight Tuning For TTS Synthesis: Reducing User Fatigue By Improving User Consistency”, In Proc. ICASSP, Toulouse, France, pp. 865-868, 2006.

[8] Vepa, J. and King, S., “Subjective evaluation of join cost functions used in unit selection speech synthesis”, In Proc INTERSPEECH, Jeju Island, Korea, pp 1181-1184. 2004.

[9] Navarro, G. and Raffinot, M., “Flexible Pattern Matching in String”, Cambridge University Press, 2002.

[10] Vepa, J., King, S. and Taylor, P., “Objective distance measures for spectral discontinuities in concatenative speech synthesis”, In ICSLP, Denver, USA, 2002.

[11] Gamboa Rosales, H., Jokisch, O. and Hoffmann, R., “Spectral distance costs for multilingual unit selection in speech synthesis”, In Proc. of 11-th International Conference “Speech and Compute” SPECOM2006, St. Petersburg, Russia, pp. 270-273,2006.

[12] Toda, T., Kawai, H., Tsuzaki, M., Shikano, K., “An Evaluation of Cost Functions Sensitively Capturing Local Degradation of Naturalness for Segment Selection in Concatenative Speech Synthesis”, Speech Communication, Vol. 48, No. 1, pp. 45-56, Jan. 2006.

[13] Hoffmann, R., “Signalanalyse und -erkennung”, Ed. Springer,1998.

[14] Gamboa Rosales, H. and Jokisch, O., “KorpusDress1 - Korpusbasierte Konkatenative Sprachsynthesesysteme”, In Proc 18. Konferenz Elektronische Sprachsignalverarbeitung, Cottbus, Germany, pp. 115-122, 2007.

[15] Coorman G., Fackrell, J., Rutten, P. and Van Coile, B.,“Segment selection in the LH Realspeak laboratory TTS system”, In Proc of ICSLP, pp. 2:395-398, Beijing, China, 2000.