Maximum Likelihood Unit Selection for Corpus-based Speech Synthesis
Abubeker Gamboa Rosales1, Hamurabi Gamboa Rosales2, Ruediger Hoffmann2

1 Engineering Department, University of Guanajuato, Mexico
2 Laboratory of Acoustics and Speech Communication, Dresden University of Technology, Germany

Abstract
Corpus-based speech synthesis systems deliver a considerable synthesis quality since the unit selection approaches have been optimized in the last decade. Unit selection attempts to find the best combination of speech unit sequences in an inventory so that the perceptual differences between expected (natural) and synthesized signals are as low as possible. However, mismatches and distortions are still possible in concatenative speech synthesis and they are normally perceptible in the synthesized waveform. Therefore, unit selection strategies and parameter tuning are still important issues in the improvement of speech synthesis. We present a novel concept to increase the efficiency of the exhaustive speech unit search within the inventory via a unit selection model. This model bases its operation on a mapping analysis of the concatenation sub-costs, a Bayes optimal classification (BOC), and a Maximum likelihood selection (MLS). The principle advantage of the proposed unit selection method is that it does not require an exhaustive training to set up weighted coefficients for target and concatenation sub-costs. It provides an alternative for unit selection but requires further optimization, e.g., by integrating target cost mapping.

Index Terms: speech synthesis, unit selection, concatenation cost, speech intelligibility, discontinuity perception.

1. Introduction
Corpus-based concatenative speech synthesis has been improved and utilized in text-to-speech synthesis (TTS) systems over the last years [1], [2]. In this approach, the speech inventory design covers a big variety of the phonetic and prosodic language features. Consequently, unit selection (US) must find the best unit sequence to synthesize an input text by minimizing the target and concatenation sub-costs. The target cost is defined as the estimation of the mismatch between a recorded acoustic speech unit and a predicted specification of the prosody modeling of the TTS system [1]. It is calculated as the weighted sum of difference between the components of the target and candidate feature vector. Likewise, concatenation cost reflects the mismatch and distortion between two speech units when they are concatenated. Target and concatenation costs are joined in a total cost function. The total cost function is described as the weighted sum of target and concatenations sub-costs, such as Duration, F0, and Energy for the target sub-costs and Linear spectral frequencies coefficients (LSFs), Multiple centroid analysis coefficients (MCAs) [3], and Mel frequency cepstral coefficients (MFCCs) for the concatenation sub-costs.

Mismatches and distortions are known as concatenation cost and they can appear because frequency formants and others spectral features between two concatenated units do not align properly [4]. Therefore, concatenation cost is considered as an good estimator of the quality of speech synthesis. Additionally, if the difference between a speech unit and the predicted specification is also taken into account, the quality of the synthesized waveform can even suffer an extra degradation. Hence, the next task is to find the weighted coefficients that determine the effect-weight of each target and concatenation sub-cost in the total cost function. This unit selection process is considered as one of the best ways 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 in a TTS System [1][5][6][7]. Here we present a unit selection method based on mapping analysis of concatenation sub-costs, Bayes optimal classification algorithm (BOC) and Maximum likelihood selection. The combination of these methods represents a novel unit selection. This unit selection method does not require an exhaustive training to set up weighted coefficients as principal advantage and principally avoids the exhaustive and subjective search for weighted coefficients. Additionally, this method estimates in great part the quality or degradation of the synthesized signal by mapping the concatenation sub-costs. In Section 2 an overview on the mapping analysis of the concatenation sub-cost used for the proposed unit selection is given. The BOC and MLS are described in section 3 and 4. Experimental results are explained in section 5.

2. Mapping Analysis
The mapping analysis consists of the (off-line) calculation of the concatenation sub-costs between speech units in the entire inventory, which do and do not present mismatches and displeasing distortions when they are concatenated. The mapping is achieved by estimating the distance between the unit features like MCAs, LSFs and MFCCs at the right and left boundaries area [8]. The concatenation sub-costs are calculated as it is described in equation (1):

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

where $C$ represents the sum of concatenation sub-costs for a sequence of $n$ speech units, $u_i$, the speech unit, $c_j^c$ concatenation sub-cost, $q$ the number of concatenation sub-costs, $w$ the weighted coefficients (WCF), and $n$ the sequence of speech units [9]. The mapping of the concatenation sub-costs that do not present any distortion is done by calculating the concatenation sub-costs between contiguous speech units in the database. The speech database is described in section 2.1. Although, the concatenation sub-costs of contiguous units are set up to zero by definition [1][2], we utilize the calculated concatenation sub-cost distances to map the real values of contiguous
speech units in the database. In this way, we obtain a real reference of concatenation sub-costs without distortion. The mapping of those concatenation sub-costs that present unpleasant distortions is done by using a determined set of speech units. These speech units were extracted from different phonetic contexts and come from different words or sentences contained in the speech database. Consequently, they do not concatenate properly in most of the cases. Therewith, the second reference is also obtained with the same number of concatenation samples like the proper concatenation samples. So, we obtain two mapped references, which represent the smooth and glitch areas of concatenation as it is shown in Fig. 1. It illustrates the mapping of the concatenation sub-cost distances between contiguous and not contiguous 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 proved to be the most inclined case to not concatenate properly [5][8].

2.1. Speech corpus and database

We utilize the “TC-STAR” English speech database [10] for the estimation of the concatenation sub-costs. The database was designed under high quality criteria. The speech quality of the recordings was reached with 96 kHz sampling rate, 16 Bits precision, SNR > 40 dB, and a bandwidth of 40 Hz to 20 kHz. The speech database has a duration of over 10 hours. It is composed of a corpus of about 90 000 words, which are contained in 5558 sentences. This amount is distributed on the sub-corpora of transcribed speech, written text, constructed phrases, and expressive speech. Seventy percent of the sentences were labeled automatically and 30% were hand labeled, where 30% cover all English diphones.

3. Bayes Optimal Classification

Once the mapping of the concatenation sub-costs was achieved, the next task is to determine the boundary area between the concatenation sub-costs that do and do not present distortion. To achieve this work, we utilize the BOC that uses a discriminant function separating the smooth and glitch areas of concatenation (off-line). Bayes optimal classification establishes that the class probability \( k \) given the feature vector \( \vec{x} \) is equal to multiplication between the a priori likelihood of the class \( P(k) \) and the density probability function \( P(\vec{x}/k) \) divided by the probability of the sample, according to equation (2):

\[
P(k/\vec{x}) = \frac{P(\vec{x}/k) \cdot P(k)}{P(\vec{x})} \tag{2}
\]

where \( k \) is the smooth or glitch 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 for contiguous and not contiguous 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 a feature vector on the class label. BOC was modeled with a multivariate Gaussian density distribution [11] considering that the feature vectors have a normal distribution as it is shown in the following equation (3):

\[
P(\vec{x}/k) = \frac{1}{(2\pi)^{N/2}|K_k|^{1/2}} \cdot \exp \left[ -\frac{1}{2}(\vec{x} - \vec{\mu}_k)^T K_k^{-1} (\vec{x} - \vec{\mu}_k) \right]
\]

where \( \vec{x} \) is the Mahalanobis distance by concatenating speech units, the covariance matrix \( K_k \), and the mean \( \vec{\mu}_k \) are calculated according to the class feature vectors of Mahalanobis distance. Because the contiguous and not contiguous concatenation sub-cost are correlated (see Fig. 1), \( K_k \) is a full covariance matrix. 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 equations (4) and (5):

\[
e = \arg \max_{i=1,...,K} d_i(\vec{x}) \tag{4}
\]

\[
d_i(\vec{x}) = P(k) \cdot P(\vec{x}/k) \tag{5}
\]

where \( e \) is the maximum argument of the discriminant function \( d_i(\vec{x}) \) in the equation (5), which contains the maximum probability. \( K \) is the number of classes (smooth or glitch concatenation type).

3.1. Bayes discriminant function

By the substitution of the multivariate Gaussian density distribution (3) in the discriminant function (4) we obtain the corresponding distance Bayes discriminant function (6). The discriminant function allows to classify a concatenation between two not contiguous speech units into glitch and smooth concatenation type, which is based on its probability estimation:

\[
d_i^*(\vec{x}) = \ln \left[ d_i(\vec{x}) (2\pi)^{N/2} \right] \tag{6}
\]

\[
= \ln P(k) - \frac{1}{2} \ln |K_k| - \frac{1}{2} \sum_{m=1}^{N} \sum_{k=1}^{K} k^{(mn)}_{m} \mu_{mk} \mu_{nk} + \sum_{m=1}^{N} \left( \sum_{n=1}^{N} k^{(mn)}_{m} \mu_{nk} \right) x_{n} - \frac{1}{2} \sum_{m=1}^{N} k^{(mn)}_{m} x_{m}^2 - \sum_{m=1}^{N-1} \sum_{n>m}^{N} k^{(mn)}_{m} x_{m} x_{n}
\]

where \( k^{(mn)}_{m} \) is the \((m, n)\)-element the inverse of \( K_k \). Equation (6) describes a Bayes discriminant function [11] and it is calculated for the corresponding glitch or smooth areas for each type.
of phoneme concatenation. The boundary between both areas can be calculated by subtracting both discriminant functions as it is shown in equation (7).

$$d_{12}(\vec{x}) = d_{3}(\vec{x}) - d_{2}(\vec{x})$$

The resulting Bayes discriminant function (7) is illustrated for a nasal phoneme /N/ concatenation type in Fig. 2. It illustrates 2-dimensional mapping analysis, where both concatenation areas are delimited by the Bayes discriminant function (7). It is easy to recognize that some concatenation sub-costs of not contiguous speech units fall inside the smooth concatenation area, which is known as classification error [11].

4. Maximum Likelihood Selection

Maximum likelihood is a well known statistical method for estimation. The concatenation sub-costs of the left over speech units are distributed normally with some unknown mean and variance. We utilized maximum likelihood method to compute the corresponding previously obtained distribution of concatenation sub-costs of the smooth concatenation type [11]. The mean of the concatenation sub-costs is then the maximum likelihood estimator of the concatenation sub-costs population, and the concatenation sub-costs variance is a close approximation to the maximum likelihood estimator of the population variance. Maximum likelihood picks the mean and variance values of the normal distribution of the concatenation sub-costs model that make the concatenations sub-costs more likely than any other values of the parameters would make them. So, maximum likelihood estimation can determine the likelihood of a bad or good concatenation between two specific units. So, the speech units that shows the highest likelihood, when they are evaluated, are selected in the unit selection process.

4.1. Unit Selection Process

Unit selection process is composed by various modules, which are shown in Fig. 3. Firstly, the speech unit candidates are chosen in the speech database by Backward Oracle Matching algorithm (BOM) [12]. BOM 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 estimated at the right and left boundaries and they are represented in a vector form. Afterwards, the distance $\Delta$ between the predecessor and candidate speech units sequence of the desired synthesized waveform is calculated. So, the speech units that had been found via BOM for the desired synthesized waveform are processed by the Bayes Classifier. The BOC classifies the speech units in glitch and smooth concatenation types by using the corresponding discriminant function, as it is shown in Fig. 2. Then, the speech units that were found to concatenate glitch and whose concatenation sub-costs do not fall into the smooth concatenation delimited area by the discriminant function, are removed from the unit selection process. So, the left over speech units are computed by the corresponding previously obtained maximum likelihood model. Finally, the speech units that show the highest likelihood, when they are concatenated, are selected. In this way, the speech units are selected that match best with the searched phoneme sequence to obtain the desired speech signal minimizing distortions in the corpus-based speech synthesis to be as low as possible.

5. Experimental Results

To evaluate the proposed unit selection method, we carried out a listening test. For the listening test the Dress TTS system was used [13]. Syrdal mentioned that a reliably higher discontinuity detection rate for diphthongs is observed than for monophthong vowels [15]. Hence, we synthesized three blocks of 10 utterances containing English diphthongs with three different unit selection methods. Also, prosody modifications were applied to the synthesized utterances to avoid jumps in the fundamental frequency and inappropriate segmental duration. The speech database “TC-STAR” was used for the three unit selection methods in the Dress TTS system. The first block was synthesized by the unit selection method proposed in [1]. This unit selection represents the basic principles of the sum of target and concatenation costs for unit selection and it requires an exhaustive training to set up the weighted coefficients for target and concatenation sub-costs. We call it Classical US. We utilized the unit selection method proposed in [14] for the second block. This method bases its operation on the previously defined transparency and quality functions so that it determines whether a determined concatenation will or will not present distortions. We call this second unit selection Masking US. The last block was synthesized with the proposed unit selection method and we called it Maximum Likelihood US.

![Figure 2: Bayes Discriminant Function.](image-url)
5.1. Listening test
The synthesized utterances were evaluated by 10 listeners. All listeners were students or researchers at Dresden University of Technology with good English proficiency and experience on speech recognition or speech synthesis. The listening test consisted of an evaluation of the intelligibility and concatenation quality of the synthesized utterances. The synthesized utterances were played in random order for each listener. We asked the listeners to rate the quality of the synthesized utterances on a scale of 1 (Bad) to 5 (Excellent). The Mean Opinion Score (MOS) values obtained for the three unit selection methods are summarized in Table 1.

Table 1: MOS Listening Test.

<table>
<thead>
<tr>
<th>Classical US</th>
<th>Masking US</th>
<th>Maximum Likelihood US</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.8</td>
<td>2.3</td>
<td>2.7</td>
</tr>
</tbody>
</table>

The mean opinion scores in Table 1 are significant on one percent-level by paired t-test. Classical US, based on the sum of target and concatenations costs, performed slightly better than Maximum Likelihood US and so much better than Masking US. This shows the potential improvements that can be obtained by taking the target sub-costs into account in the unit selection process. Nevertheless, the task of setting up the weighted coefficients on the total cost function is a very difficult subjective work, which requires many hours of listening training for the specific corpus database [7]. The proposed Maximum Likelihood US obtained better results than the Masking US and it was slightly worse than the Classical US manifesting only a small perceptive difference between them. Masking US 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 each type of concatenation as it was proposed by [14].

6. Conclusion
Summarizing, this paper presented another perspective on unit selection methods for corpus-based speech synthesis by proposing a Bayes optimal classifier with a Maximum Likelihood selection. The proposed unit selection method is based on concatenation sub-costs mapping of speech units in the database. The mapping provides two references of smooth and glitch concatenation areas. Furthermore, a discriminant function as shown in equation (7) was developed, which calculates the probability estimation of smooth and glitch concatenation type between two speech units. Finally, a maximum likelihood method is utilized to select the speech units that are characterized as smooth, when they are concatenated. Maximum likelihood unit selection has one principle advantage because it does not require an exhaustive training to set up the weighted coefficients for target and concatenation sub-costs. This method bases its operation on an objective mapping of the concatenation sub-costs. Therefore, maximum likelihood unit selection supports the integration of new speech databases in a TTS system avoiding exhaustive training for each newly integrated speech database. Additionally, the proposed unit selection method delivers a functional performance with an acceptable quality of speech synthesis.

In future, it will be important to improve the proposed unit selection performance by the integration of a target sub-costs mapping, because the target sub-costs have shown a great influence on the intelligibility and naturalness of the synthesized speech signal.

7. References