University of West Bohemia in Pilsen

Faculty of Applied Sciences

Dissertation

2012 Ing. Milan Legát

University of West Bohemia in Pilsen
Faculty of Applied Sciences

CONCATENATION COST IN UNIT SELECTION SPEECH SYNTHESIS

Ing. Milan Legát

Thesis submitted for the degree of
Doctor of Philosophy in the field of
Cybernetics

Supervisor: Doc. Ing. Jindřich Matoušek, Ph.D.
Department: Department of Cybernetics

Pilsen 2012

Západočeská univerzita v Plzni
Fakulta aplikovaných věd

CENA ŘETĚZENÍ V SYNTÉZE ŘEČI VÝBĚREM JEDNOTEK

Ing. Milan Legát

disertační práce

k získání akademického titulu doktor
v oboru Kybernetika

Školitel: Doc. Ing. Jindřich Matoušek, Ph.D.

Katedra: Katedra Kybernetiky

Plzeň 2012

Dedicated

To
My Family and Veronika

Acknowledgments

I want to thank to all my colleagues from the TTS team at the Department of Cybernetics at the Faculty of Applied Sciences, University of West Bohemia in Pilsen. Firstly, I thank to my supervisor Jindřich Matoušek who has guided me throughout my work by referring me to interesting papers and book chapters, and also by inspiring consultations. Special thanks belong to Jan Romportl for providing us with a great online interface for conducting listening tests. I also thank to Daniel Tihelka for creating an experimental version of the ARTIC TTS system and describing the current implementation of the system. I thank to Martin Grůber who I worked with on the R library for training classification and regression models.

I also want to thank to my colleagues from the former research team of SVOX—Johan Wouters, Alexis Wilpert, Igor Nozhov, Marc Padellini, Joram Meron and Gert Coorman. The numerous discussions we had, albeit not directly related to the work on this thesis, were a valuable source of ideas that inspired my work in many ways.

I thank to my colleague from the Nuance TTS Innovation Team, Virginie Holm, for translating the résumé of this work into French.

Big thanks belong to my family. Without their endless support I would have never got where I am now. Finally, I thank to my fiancée, Veronika, for her patience over the many years I have been spending weekends and long evenings working on this thesis.

Declaration

I declare that, apart from where properly indicated, this thesis is the result of my own original work.

In Pilsen November 30, 2012 Milan Legát

Abstract

This thesis deals with one of the key aspects of the unit selection speech synthesis method—**design of a concatenation cost function**. The concatenation cost function measures quality of concatenations of units that are taken from a unit database at synthesis runtime. Ideally, the obtained measurements should correlate well with human perception of the quality of the concatenations.

Design of a concatenation cost function is a complex problem due to a wide range of different audible artifacts that can be encountered at concatenation points. Therefore, the scope of the work is narrowed to five short Czech vowels and two speakers—one female and one male.

In the first part of the work, a method for collecting reliably annotated data have been proposed. It is shown that the method allows for obtaining well correlated annotations of the quality of concatenation points. At the same time, the data are rich enough for measuring different sources of discontinuities.

It is generally believed that concatenation discontinuities have to be measured at least in three dimensions—energy, $F0$ and spectrum [Dut08]. This work mainly investigates the role of $F0$, which is found to be crucial for the quality of mid-vowel concatenations, and the role of consonantal contexts that can change the spectral content of concatenated vowel instances as a result of coarticulation.

It is shown that the consonantal contexts have only a limited impact on the quality of the concatenations, in contrast to $F0$. The $F0$ discontinuities however have to be measured by using $F0$ contours capturing the dynamics of $F0$ in concatenation areas rather than by calculating static $F0$ differences at concatenation points, which is the traditional approach. Information contained in the $F0$ contours is sufficient for explaining a vast majority of audible discontinuities.

The work on the concatenation cost is put into a wider perspective of the unit selection method by proposing an analytic method that allows for measuring the perceptual relevance of different costs, cost sub-components and their weights.

Contents

List of Figures

List of Tables

Notation and List of Symbols

a	a scalar variable
$\mathbf{a}$	a column vector or a vector of parameters
$\mathbf{A}$	a matrix
$\|\|\mathbf{a}\|\|$	the Euclidean norm of vector $\mathbf{a}$
$\|\mathbf{A}\|$	the determinant of matrix $\mathbf{A}$
$\mathbf{A}^T$	the transpose of matrix $\mathbf{A}$
$\hat{a}$	an estimate of parameter a
t	indexes time
n	indexes signal samples

All phonetic transcriptions presented throughout this work are using the Czech SAMPA symbols. Slash symbol denotes starts and ends of phonetic transcriptions.

Terms that are introduced in this work are written in *Italics*.

Chapter 1

Introduction

1.1 Preface

The goal of this chapter is to provide the reader with sufficient background information needed for understanding the topic of this thesis. We will start with a general introduction into the field of speech synthesis. From this general introduction, we will move on to more specific information about concatenation cost functions with the primary focus on the state-of-the-art of their design. Finally, the objectives and the structure of this thesis will be presented.

1.2 Speech Synthesis

Speech synthesis is, generally speaking, a process of production of artificial speech. Its history dates back to the end of the 18th century, when Wolfgang von Kempelen constructed his first mechanical talking machine. Since that time, the development within the field of speech synthesis has made a huge progress, and especially the intelligibility of the current state-of-the-art systems reaches a high level.

Throughout more than two centuries various methods have been proposed including electronic speech synthesizers, formant, concatenative, sinwave, HMM-based and articulatory synthesis. In addition, some hybrid methods have also been introduced.

1.3 Unit Selection-based Concatenative Speech Synthesis

Despite the increasing popularity of HMM-based speech synthesis, the unit selection concatenative systems still represent the mainstream in many practical applications, especially in limited domains where synthesized chunks are combined with pre-recorded prompts. In such applications, the ability of the unit selection to deliver highly natural and to the recordings well fitting output is the key factor. Not surprisingly, the unit selection also remains the first choice for eBook reading applications, which have been acquiring a lot of interest over recent years.

1.3.1 What is Unit Selection?

Unit selection speech synthesis systems generate their output by selecting and concatenating in some sense optimal sequences of units that are taken from a large inventory. These units are basically chunks of segmented recordings, and their size may vary from sub-phoneme units (typically half phones) to multiple phoneme units (e.g. syllables). The units may also be of variable sizes.

Since each target (e.g. a diphone) is typically represented by more than one unit in the inventory, a selection mechanism using two cost functions—*target cost* and *concatenation (join) cost*—is applied to find the best sequence of units [HB96]. This mechanism usually works as an optimal path search routine operating within a lattice of units—*Viterbi search*. Each node of the lattice represents a single instance of a unit and is given a target cost value. Transitions between the nodes are assigned concatenation cost values (see Fig. 1.1).

It is obviously not feasible in the open domain to cover all possible inputs with continuous sequences of units in the inventory. If there is however a reliable mechanism for selecting units which sound well when concatenated, it is possible, thanks to the variety in acoustic characteristics of the units, to generate highly natural and intelligible speech with minimum or even no smoothing at the concatenation points required. This assumption is sometimes referred to as *take the best to modify the least* [BPQ$^+$99].

1.3.2 Target Cost Function

General TTS system can normally be divided into two parts—*front-end* and *back-end* [Sch06]. The front-end is a part of the system that is closer to an input text. Its output is typically phonetic transcription of the input text enriched with additional information such as prosodic features, word stress, phrasing, etc. This representation specifies a sequence of target units to be synthesized.

The target cost function estimates a perceptual distance between a target unit and theoretically every unit in the inventory of a system. This cost function typically consists of several subcomponents representing a combination of *phonological features* such as an identity of the unit and/or its context, *positional features* and *numerical features* such as duration, fundamental frequency[1] ($F0$) or intensity of the given unit [Tih05b].

More formally, the target cost, $C^t\left(t_i, u_i\right)$, can be calculated as a sum of weighted feature vector differences between a candidate, u_i, and the target, t_i, units [HB96]:

$$C^t\left(t_i, u_i\right) = \sum_{j=1}^{p} w_j^t C_j^t\left(t_i, u_i\right), \tag{1.1}$$

[1] The fundamental glottal frequency, $F0$, is the frequency of the cords oscillations and characterizes the fundamental tone of human voice.

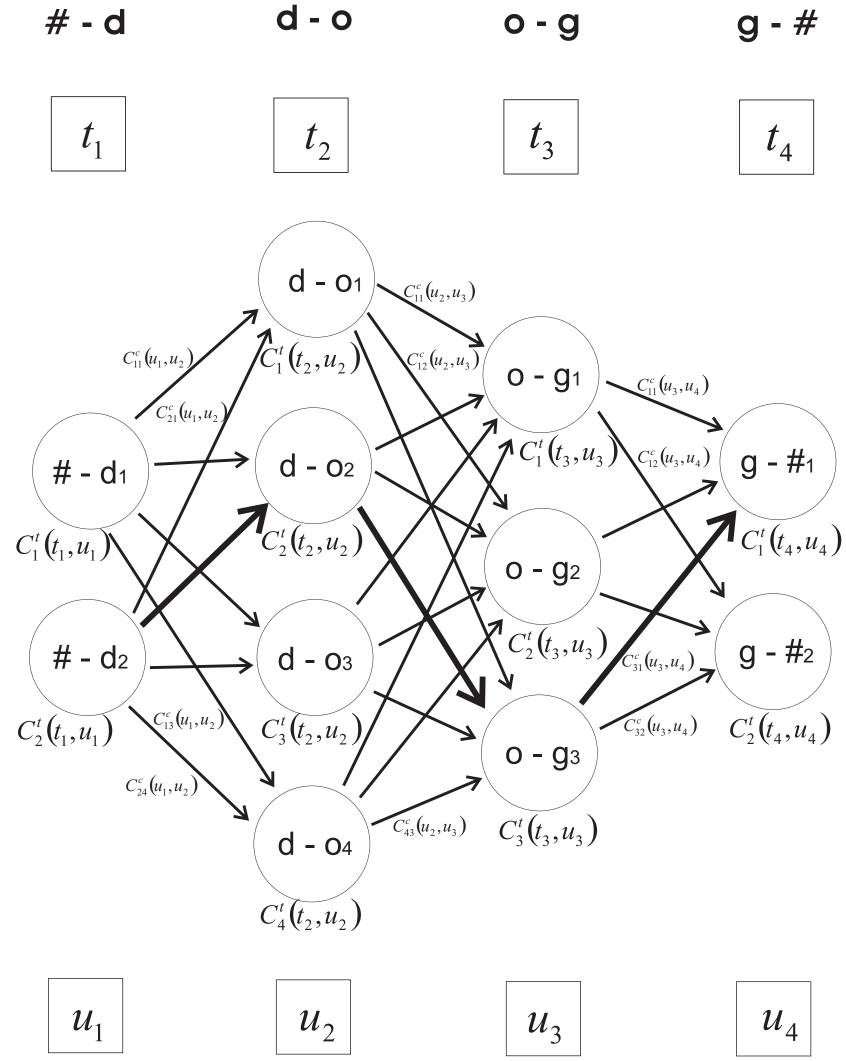


Fig. 1.1: Demonstration of the unit selection method

Each node of the state transition network represents a single instance of a unit (e.g. a diphone) and is given a value of the target cost function, $C^t\left(t_i, u_i\right)$. The transitions between units are denoted by values of the concatenation cost function, $C^c\left(u_{i-1}, u_i\right)$ (not all concatenation costs are depicted in the figure, so as to maintain its clarity). In order to synthesize the word "dog", the Viterbi algorithm is applied to find the path with a minimum overall cost through the lattice (emphasized with thick arrows). Note that a real unit inventory contains many more instances than shown in this figure.

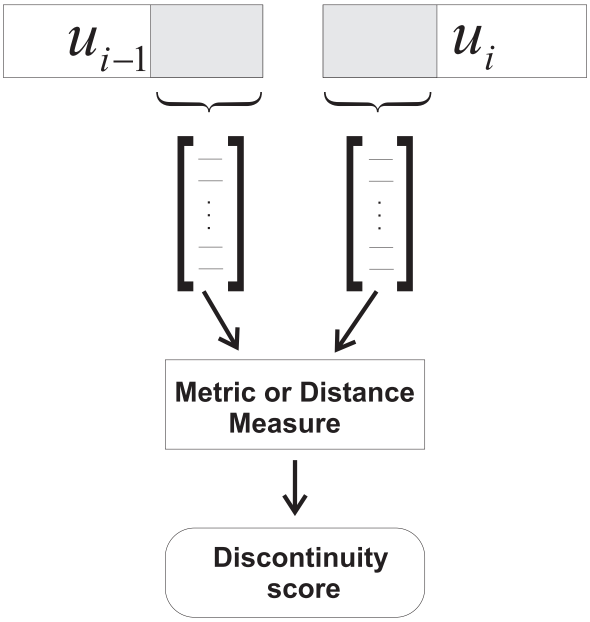


Fig. 1.2: Traditional calculation of the concatenation cost

Feature vectors are extracted from adjacent segments of concatenated units, then a metric or distance measure is applied. The result is a discontinuity score which should ideally correlate well with a perceived level of discontinuity.

where p represents the number of the target cost components, and w_j^t is a feature weight of the j-th component.

1.3.3 Concatenation Cost Function

Traditional implementations of the concatenation cost functions are exhaustively addressed in Sec. 1.4. This section only serves as a brief introduction to complete the description of the unit selection method.

While the task of the target cost function is to estimate a perceptual difference between the target and a candidate unit, the concatenation cost function should reflect a level of perceived discontinuity between two concatenated units. It consists, similarly to the target cost function, of a set of subcomponents which may be associated to a difference in pitch, energy and spectra of adjacent segments of the concatenated units (see Fig. 1.2), but it can have a phonological nature as well. The concatenation cost function can also be defined in a way to reflect trajectories of different features across larger neighbourhood of concatenation points. The concatenation cost is traditionally calculated as follows [HB96]:

$$C^c\left(u_{i-1},u_i\right)=\sum_{j=1}^{q} w_j^c C_j^c\left(u_{i-1},u_i\right), \tag{1.2}$$

where q represents the number of the concatenation cost components, and w_j^c is a feature weight of the j-th component.

In the special case, when u_{i-1} and u_i are two units which stand next to each other in the unit inventory, i.e. there is a natural transition between them, the concatenation cost is set to zero. This encourages the selection of larger sequences of originally consecutive units, which leads to high naturalness of synthesized speech. The selection of units with very low concatenation cost and high target cost may however lead to lower intelligibility [Dut08]. Thus, these two cost functions need to be properly balanced.

1.3.4 Searching for Optimal Sequence of Units

To choose the optimal sequence of units $\overline{u}_1^n$ from the unit inventory, the sum of the target and the concatenation cost functions is minimized [HB96]:

$$\overline{u}_1^n=\min_{u_1,\ldots,u_n} C\left(t_1^n,u_1^n\right), \tag{1.3}$$

where

$$C\left(t_1^n,u_1^n\right)=\sum_{i=1}^{n} C^t\left(t_i,u_i\right)+\sum_{i=2}^{n} C^c\left(u_{i-1},u_i\right)+C^c\left(S,u_1\right)+C^c\left(u_n,S\right), \tag{1.4}$$

and S denotes silence.

The minimization of this sum can also be interpreted as a search for the optimal path through a state transition network (see Fig. 1.1), for which the Viterbi search [Vit67] is typically employed.

Different pruning methods are often required to speed up the synthesis on large speech inventories that may contain tens of thousands of unit candidates [HB96]. In addition or instead of pruning, various clustering methods can also be applied to lower the computational requirements of the calculation of the target costs [BT97], [Oli77]. Regarding the computation of the concatenation costs, which is by far the most time-consuming, a limited number of unit pairs' concatenation costs can be pre-calculated and stored in memory [BMR99].

1.4 Concatenation Cost Functions

1.4.1 Overview

As already mentioned in Sec. 1.3.3, the concatenation cost function is one of the two cost functions that are typically used when searching for the optimal sequence of units in the unit selection speech synthesis. Its task is to measure discontinuities which may appear as a result of concatenating neighbouring units. The ideal concatenation cost function should correlate well with the human perception of the concatenation discontinuities.

Many studies dealing with the concatenation cost design have been published over the last one and a half decades. These studies have primarily been focusing on spectral mismatches while eliminating other possible sources of discontinuities. An effort has typically been made to find the right parametrization of the spectrum of speech signals, and combine it with an appropriate distance measure (e.g. the Euclidean distance, the Mahalanobis distance or the symmetrical Kullback-Leibler distance, to name but a few).

The parameterizations which have been experimented with include, but are not limited to, mel-frequency cepstral coefficients (MFCCs) [BSF05], [KOS06a], [SS01], linear prediction coefficients (LPCs) [BSF05], [KOS06a], [KV01], linear prediction cepstral coefficients (LPCCs) [KV01], line-spectrum frequencies (LSFs) [KOSE07b], [PS07], [SS01], [VK04], residual MFCCs, bispectrum [PS07], [CC99], modified Mellin transforms of the log spectrum (MMTLS), Wigner-Ville distribution-based cepstrum (WVD) [CC99], FFT-based cepstra, log area ratios (LAR), perceptual linear prediction coefficients (PLPs), multiple-centroid analysis coefficients (MCAs) [VK04] or formant frequencies [KV01].

In addition to the traditional approaches mentioned above, some other alternative techniques have been proposed. These, for example, include the application of the Latent Semantic Mapping (LSM) [Bel04], Kalman filters [VK03], or the Auditory Image Modelling [Tsu01], [TK02]. Also phonetic features have been proposed as an option [KT02]. In the following subsections, some of these non-traditional concatenation cost functions will be described in more detail.

1.4.2 Kalman Filter-based Concatenation Cost

Introduction

In [VK03], a concatenation cost function based on Kalman filtering was proposed. The reader can also refer to [Vep04] for more details.

The proposed method is based on the utilization of Kalman filters (linear dynamical models) to learn an underlying representation of speech signals—line spectral frequencies (LSF) in this case. The models were trained on natural speech separately for each phoneme and the concatenation cost was then derived from a difference between predicted and observed trajectories of the LSFs at the concatenation points (see Fig. 1.3).

Design and Training of Kalman Filter

The filters presented in the above mentioned works have been implemented according to [Fra03], and can be described using the following equations:

$$\mathbf{y}_t = \mathbf{H}\mathbf{x}_t + \epsilon_t \qquad \epsilon_t \sim N\left(\mathbf{v}, \mathbf{C}\right) \tag{1.5}$$

$$\mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \eta_t \qquad \eta_t \sim N\left(\mathbf{w}, \mathbf{D}\right), \tag{1.6}$$

where $\mathbf{y}_t$ is an observed feature vector (LSF in this case), $\mathbf{x}_t$ is a state vector starting from the initial condition $\mathbf{x}_0 \sim N\left(\pi, \mathbf{\Lambda}\right)$ and ϵ_t, η_t are uncorrelated normally distributed noise vectors with means $\mathbf{v}$, $\mathbf{w}$ and covariance matrices $\mathbf{C}$, $\mathbf{D}$, respectively.

Before the models can be used, their parameters ($\mathbf{H}$, $\mathbf{F}$, $\mathbf{v}$, $\mathbf{w}$, $\mathbf{C}$, $\mathbf{D}$, $\mathbf{x}_0$) need to be found. This was accomplished by employing the Expectation Maximization (EM) algorithm [DLR77] using natural speech (LSFs extracted from natural speech, more precisely). Prior to the EM algorithm, the model parameters had been initialized using three different initialization schemes—the first order autoregressive process modelling (similar to [GH96]), factor analysis [Fra03], [RG99], and hand-picking and tuning of the initial parameters based on experiments done by [Fra03].

Concatenation Cost Design

One set of parameters of the Kalman filter was found for each phoneme. It was assumed that the model trained on the LSF trajectories of natural speech is capable to follow these trajectories precisely at the beginning and at the end of a phoneme as these are the natural segments. At a concatenation point, where some discontinuity may appear, the model predicts the most likely path through the join, which may not be in accordance with the actual observation. It was this difference in predicted and actual values of the LSFs which formed the basis of the proposed concatenation cost function.

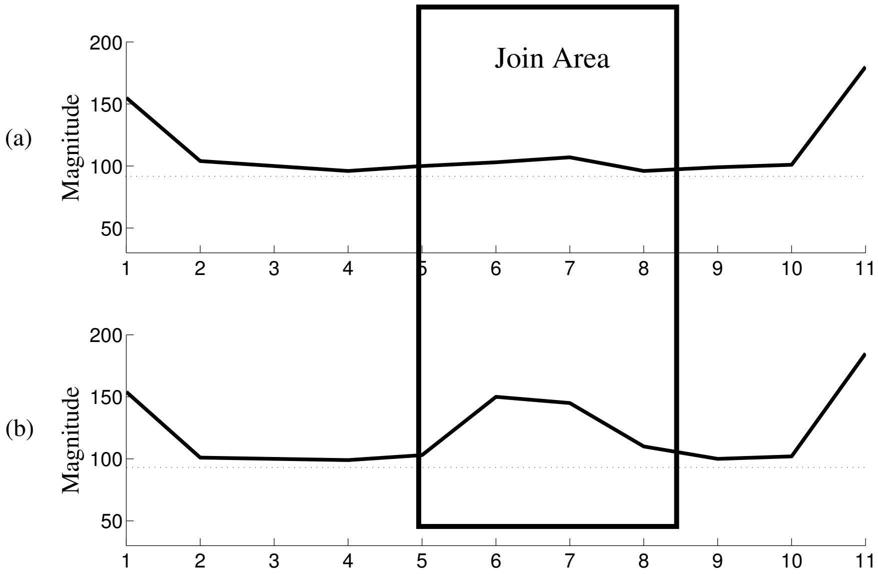


Fig. 1.3: Comparison of log likelihood estimates for a bad and a natural join (after [Vep04]).

> *(a) Negative log likelihood estimate derived from LSF trajectories (predicted and actual) for a natural join. (b) Negative log likelihood estimate derived from LSF trajectories (predicted and actual) for a synthetic join.*

For measuring of the difference between a natural and a synthetic phoneme, first the log likelihood of an observation sequence $\mathbf{Y}_{synt}$ given the parameters of the model m was calculated according to [DRO93]:

$$\log p\left(\mathbf{Y}_{synt}|m\right) = -\sum_{t=t_{start}}^{t=t_{end}} \left\{\log|\Sigma_{\mathbf{e}_t}| + \mathbf{e}_t^T\Sigma_{\mathbf{e}_t}^{-1}\mathbf{e}_t\right\} + const., \tag{1.7}$$

where $\mathbf{e}_t$ and $\Sigma_{\mathbf{e}_t}$ are the prediction error of the model m and its covariance obtained by the standard Kalman filter recursion. In Fig. 1.3, an example of the negative log likelihood estimate for a bad and a natural join are depicted.

The Kalman filter prediction error accumulated at the concatenation area due to the unnatural course of the LSF trajectories appears in the plot of the log likelihood as a “lobe”. The objective measure of discontinuity was derived from the area under this “lobe”. It was assumed that the larger the area under the “lobe” is, the more likely a particular join is perceived as discontinuous.

Evaluation

The proposed concatenation cost function was tested on a set of concatenation points in the middle of five American English diphthongs — /eI/, /oU/, /aI/, /aU/ and /oI/(in SAMPA notation). A set of sentences was synthesized, in which only diphone candidates that form the given diphthongs were altered while the rest of the sentences was kept unchanged. Sentences with poor joins in close neighbourhood of the diphthongs of the interest were removed. Further pruning based on values of a target cost function was then also applied resulting in a set of sentences containing 30 items. A listening test, in which 17 mostly native British English listeners with some experience in speech synthesis took part, was conducted. The quality of the joins was rated on a 5-point scale. A validation set of sentences was added to the listening test stimuli allowing to measure the consistency of the listeners.

The correlations between perceptual scores (mean listener scores) and scores returned by the proposed concatenation cost function were calculated. The results were compared with results obtained using concatenation cost functions based on Mahalanobis distance between MFCCs, LSFs and their deltas, absolute distance between MCA parameters and their deltas, and also absolute distance between weighted MCA parameters (refer to [VKT02b], [VKT02a] for these measures, and [VK03] for the detailed comparison of the results).

The results have shown that the Kalman-filter based concatenation cost function performs better than cost functions based on the Mahalanobis distance between MFCCs and LSFs, similarly to the cost function using the MCA parameters, but worse than the cost function based on the absolute distance between weighted MCA parameters. It has also shown that none of the evaluated concatenation cost functions performs well in all cases.

The authors however believed that there was still a room for improvement using this approach, and suggested training the models using articulatory data instead of a spectral parameterization.

1.4.3 LSM-based Concatenation Cost

Introduction

As mentioned in Sec. 1.3.3, most of the proposed concatenation cost functions rely on various spectrum-derived feature representations based on the standard Fourier analysis. Typically, only magnitude spectrum is used, whereas phase information is discarded. According to [Bel04], this may be the reason why none of such methods achieve satisfactory correlation with human perception.

In [Bel04], an alternative approach to the concatenation cost function design was introduced. This approach is based on methods of latent semantic analysis (LSA), which was originally formulated as a tool for information retrieval. The LSM[2] is built upon the assumption that there is an underlying latent semantic structure in data obscured by a noise, and that this structure can be found using algebraic and/or statistical techniques [Bel05].

The idea of exploiting the LSM paradigm for the design of a concatenation cost function is based on a matrix-style modal analysis performed via singular value decomposition (SVD). Rows of the matrix being analyzed are composed of speech data collected pitch-synchronously in a vicinity of a concatenation point and the discontinuity metric is then defined as a distance between vectors representing the concatenation area in a derived vector space of lower dimension.

[2] There are some fairly generic properties of LSA which allowed its utilization in various areas, not always directly language related, and led to the change of terminology to the more generic “latent semantic mapping” (LSM) [Bel05].

Singular Value Decomposition and Closeness Measure

Singular value decomposition is a technique which is closely related to the eigenvector decomposition and the factor analysis [CW85]. The R-order decomposition of a matrix $\mathbf{W}$ $(M \times N)$ is performed as follows:

$$\mathbf{W} \approx \hat{\mathbf{W}} = \mathbf{U}\mathbf{S}\mathbf{V}^T, \tag{1.8}$$

where $\mathbf{U}$ is the $(M \times R)$ left singular matrix with row vectors $\mathbf{u}_i$ $(1 \leq i \leq M)$, $\mathbf{S}$ is the $(R \times R)$ diagonal matrix of singular values $s_1 \geq s_2 \geq \cdots \geq s_R > 0$, $\mathbf{V}$ is the $(N \times R)$ singular matrix with row vectors $\mathbf{v}_j \, (1 \leq j \leq N)$, and $R < \min(M, N)$ is the order of the decomposition.

Both left and right singular matrices $\mathbf{U}$ and $\mathbf{V}$ are column orthonormal[3], and thus, their column vectors define an orthonormal basis for the vector space of dimension R, referred to as the LSM space [Bel05]. From (1.8), each row $\mathbf{c}_i$ of $\mathbf{W}$ can be expressed as:

$$\mathbf{c}_i = \mathbf{u}_i\mathbf{S}\mathbf{V}^T = \bar{\mathbf{u}}_i\mathbf{V}^T, \tag{1.9}$$

and after post–multiplying by $\mathbf{V}$, we obtain:

$$\bar{\mathbf{u}}_i = \mathbf{u}_i\mathbf{S} = \mathbf{c}_i\mathbf{V}. \tag{1.10}$$

The inner product of $\mathbf{c}_i$ with the k-th right singular vector can be interpreted as a measure of a strength of a signal at the mode represented by this right singular vector, and the corresponding element of $\bar{\mathbf{u}}_i$ is a real-valued coefficient of a projection of $\mathbf{c}_i$ onto this particular right singular vector [Bel04].

To compare two vectors $\bar{\mathbf{u}}_k$ and $\bar{\mathbf{u}}_l$ in the SVD-derived feature space, cosine

[3] $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I_R}$

of the angle between these vectors can be used [Bel00]:

$$K\left(\bar{\mathbf{u}}_k, \bar{\mathbf{u}}_l\right) = \cos\left(\mathbf{u}_k\mathbf{S}, \mathbf{u}_l\mathbf{S}\right) = \frac{\mathbf{u}_k\mathbf{S}^2\mathbf{u}_l^T}{\left\|\mathbf{u}_k\mathbf{S}\right\| \left\|\mathbf{u}_l\mathbf{S}\right\|}. \tag{1.11}$$

Based on this measure, various distance metrics can be designed.

Concatenation Cost Design

When designing a concatenation cost function based on the LSM, the first step is to create an input matrix to be decomposed. In the following paragraphs, we describe how this input matrix can be created according to [Bel04].

Let us consider the diphone-style concatenations, and denote P a phoneme to be synthesized. Let S_1 be a trailing speech segment of the left half of P, and S_2 a speech segment which starts with the right half of P. Let R_1 and L_2 be segments contiguous to S_1 on the right and to S_2 on the left, respectively. Further, assume that the phoneme P is synthesized as $S_1 - S_2$ but in the database only segments $S_1 - R_1$ and $L_2 - S_2$ are available.

Let us further denote $\mathbf{p}_K \dots \mathbf{p}_1$ the last K pitch periods of the signal of S_1, and $\bar{\mathbf{p}}_1 \dots \bar{\mathbf{p}}_K$ the first K pitch periods of the signal of R_1. Similarly, $\mathbf{q}_1 \dots \mathbf{q}_K$ denotes the first K pitch periods of the signal of S_2, and $\bar{\mathbf{q}}_K \dots \bar{\mathbf{q}}_1$ the last K pitch periods of the signal of L_2. Thus, when we put segments S_1 and S_2 together, the boundary region is spanned by $\mathbf{p}_K \dots \mathbf{p}_1\ \mathbf{q}_1 \dots \mathbf{q}_K$, and the concatenation point falls in the middle of this region.

The samples of $\mathbf{p}_K \dots \mathbf{p}_1\ \bar{\mathbf{p}}_1 \dots \bar{\mathbf{p}}_K$ can be moved along the signal resulting in the span $\pi_{-K+1} \dots \pi_0 \dots \pi_{K-1}$, where π_0 is composed of the right half of $\mathbf{p}_1$ and the left half of $\bar{\mathbf{p}}_1$, π_{-k} comprises the right half of $\mathbf{p}_{k+1}$ and the left half of $\mathbf{p}_k$, and π_k consists of the right half of $\bar{\mathbf{p}}_k$ and the left half of $\bar{\mathbf{p}}_{k+1}$ for $1 \leq k \leq K-1$. In other words, we obtain $2K-1$ pitch periods of the signal with the boundary falling exactly in the middle of π_0. The same operation can be applied on $\mathbf{q}_1 \dots \mathbf{q}_K\ \bar{\mathbf{q}}_K \dots \bar{\mathbf{q}}_1$, resulting in $\sigma_{-K+1} \dots \sigma_0 \dots \sigma_{K-1}$.

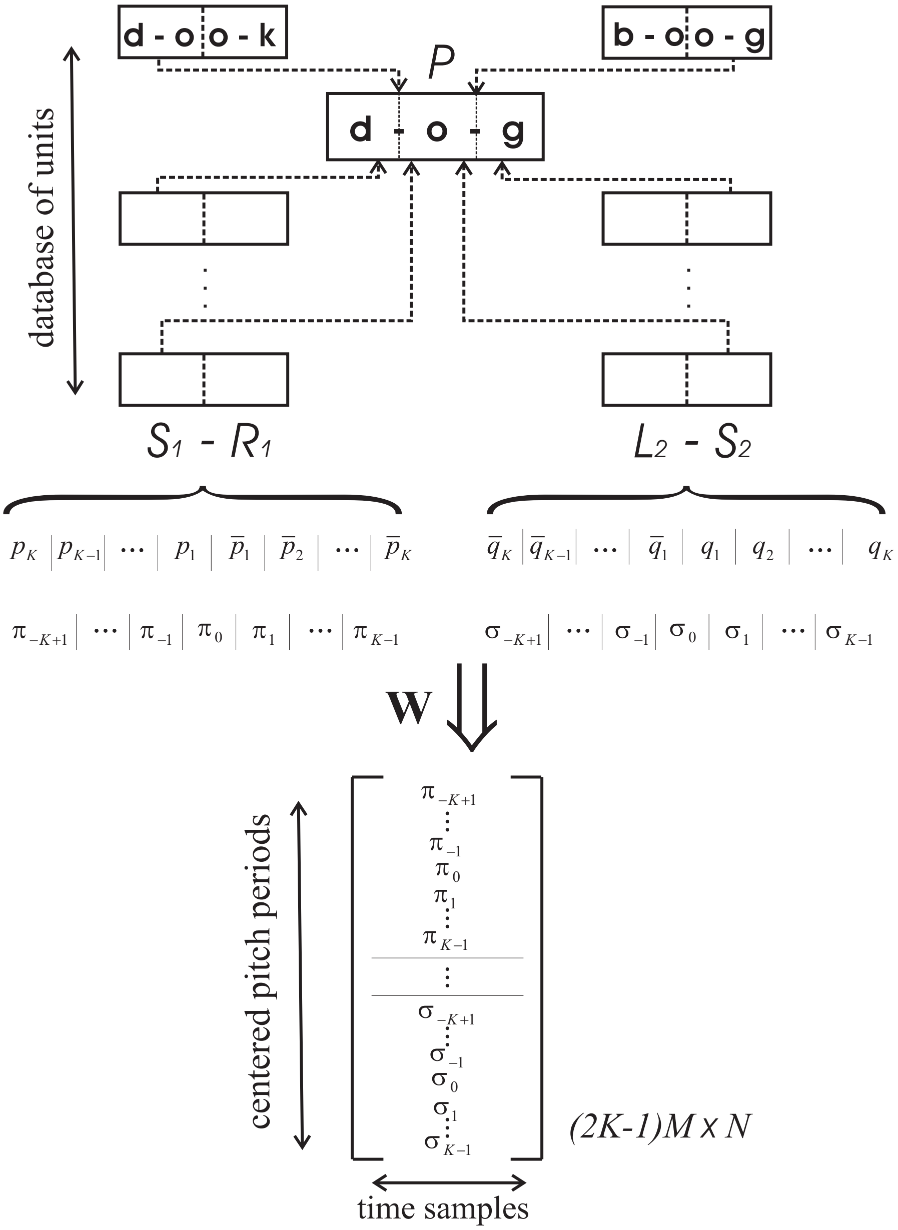

d - o o - k
b - o o - g
P
d - o - g
database of units
S1 - R1
L2 - S2
p_K p_{K-1} ⋯ p_1 $\bar{p}_1$ $\bar{p}_2$ ⋯ $\bar{p}_K$
$\bar{q}_K$ $\bar{q}_{K-1}$ ⋯ $\bar{q}_1$ q_1 q_2 ⋯ q_K
π_{-K+1} ⋯ π_{-1} π_0 π_1 ⋯ π_{K-1}
σ_{-K+1} ⋯ σ_{-1} σ_0 σ_1 ⋯ σ_{K-1}
W
centered pitch periods
π_{-K+1}
π_{-1}
π_0
π_1
π_{K-1}
σ_{-K+1}
σ_{-1}
σ_0
σ_1
σ_{K-1}
(2K-1)M x N
time samples

Fig. 1.4: Construction of an input matrix for the SVD

At the beginning, we have speech segments $S_i - R_i$ and $L_j - S_j$ from which a phoneme P can be synthesized. We extract the last K pitch periods from S_i, and K initial pitch periods from S_j. These pitch periods are then centered and form rows of a matrix $\boldsymbol{W}$ which is the input for the SVD.

Let M be the number of segments $S_1 - R_1$ and $L_2 - S_2$ present in the database, i.e. units from which the phoneme P can be synthesized. If the same pitch period extraction as described above is applied on each of these segments, we obtain $(2K-1)\,M$ centered pitch periods in total[4]. Assuming that N is the maximum number of samples within a pitch period, we obtain a $((2K-1)\,M \times N)$ matrix $\mathbf{W}$ with elements $w_{i,j}$, where each row $\mathbf{c}_i$ corresponds to a centered pitch period, and each column $\mathbf{t}_j$ to a slice of time samples. The whole process of the creation of the input matrix is shown in Fig. 1.4.

Having the input matrix, the SVD can be performed according to (1.8). The decomposition defines an alternative space for mapping between centered pitch periods (rows $\mathbf{c}_i$ of the input matrix) and feature vectors $\bar{\mathbf{u}}_i = \mathbf{u}_i\mathbf{S}$ (see Fig. 1.5). The concatenation cost function is then built upon the assumption that the "closer" two feature vectors $\bar{\mathbf{u}}_k$ and $\bar{\mathbf{u}}_l$ are in the alternative space, the less discontinuity is expected to be audible at the concatenation point.

As mentioned above, the segment $S_1 - S_2$, which is not available in the speech inventory, can be synthesized using the left half of $S_1 - R_1$ and the right half of $L_2 - S_2$, respectively. The resulting concatenation area can then be described as $\pi_{-K+1} \ldots \pi_1\ \delta_0\ \sigma_1 \ldots \sigma_{K-1}$, where δ_0 is the concatenated centered pitch period, and consists of samples of the left half of π_0 and samples of the right half of σ_0. The representation of this sequence in the LSM space can be described as follows:

$$\bar{\mathbf{u}}_{\pi_{-K+1}} \ldots \bar{\mathbf{u}}_{\pi_1}\ \bar{\mathbf{u}}_{\delta_0}\ \bar{\mathbf{u}}_{\sigma_1} \ldots \bar{\mathbf{u}}_{\sigma_{K-1}}. \tag{1.12}$$

It is obvious that the only vector which does not have a corresponding row in the original input matrix $\mathbf{W}$ is $\bar{\mathbf{u}}_{\delta_0}$. It can, however, be treated as an

[4] The recommended setting of K is $K = 3$ [Bel04].

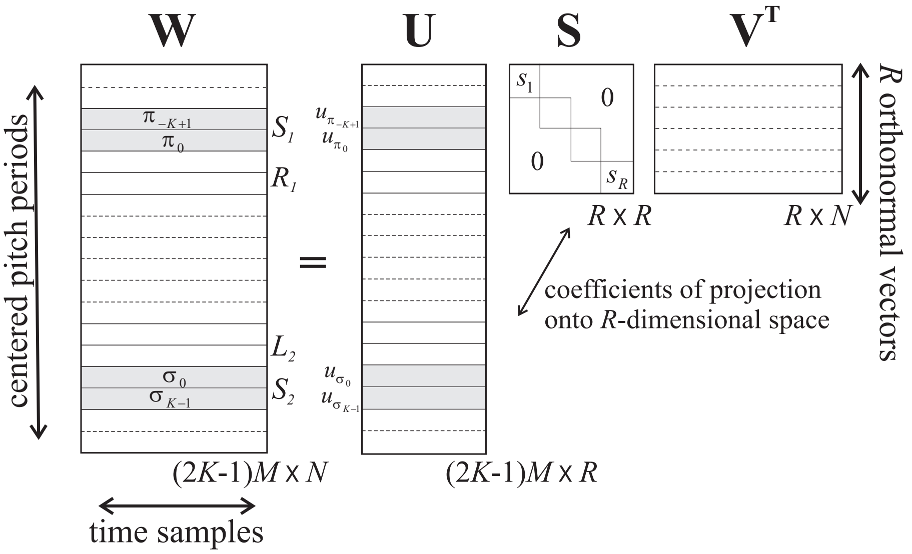


Fig. 1.5: Decomposition of an input matrix (after [Bel04])

SVD decomposition is applied on an input matrix $\mathbf{W}$*, resulting in an alternative representation of centered pitch periods (*$\bar{\mathbf{u}}_i = \mathbf{u}_i S$*).*

additional row of this matrix, and using (1.10), we obtain:

$$\bar{\mathbf{u}}_{\delta_0} = \mathbf{u}_{\delta_0}\mathbf{S} = \delta_0\mathbf{V}. \tag{1.13}$$

To measure a level of discontinuity at the concatenation point in the middle of $S_1 - S_2$, in [Bel04], the distance based on the measure (1.11) has been defined using the shorthand notation:

$$\tilde{K}\left(\mathbf{u}_{\sigma_{-1}}, \mathbf{u}_{\sigma_0}, \mathbf{u}_{\sigma_1}\right) = \frac{K\left(\bar{\mathbf{u}}_{\sigma_{-1}}, \bar{\mathbf{u}}_{\sigma_0}\right) + K\left(\bar{\mathbf{u}}_{\sigma_0}, \bar{\mathbf{u}}_{\sigma_1}\right)}{2} \tag{1.14}$$

as follows:

$$d\left(S_1, S_2\right) = 2\tilde{K}\left(\mathbf{u}_{\pi_1}, \mathbf{u}_{\delta_0}, \mathbf{u}_{\sigma_1}\right) - \tilde{K}\left(\mathbf{u}_{\pi_1}, \mathbf{u}_{\pi_0}, \mathbf{u}_{\pi_{-1}}\right) - \tilde{K}\left(\mathbf{u}_{\sigma_{-1}}, \mathbf{u}_{\sigma_0}, \mathbf{u}_{\sigma_1}\right). \tag{1.15}$$

To obtain a broader insight to the concatenation area, the distance measure (1.15) can simply be generalized. Using the following expression:

$$d\left(S_1, S_2\right) = \sum_{k=1}^{K-1} \left\{ 2\tilde{K}\left(\mathbf{u}_{\pi_k}, \mathbf{u}_{\delta_0}, \mathbf{u}_{\sigma_k}\right) - \tilde{K}\left(\mathbf{u}_{\pi_k}, \mathbf{u}_{\pi_0}, \mathbf{u}_{\pi_{-k}}\right) - \tilde{K}\left(\mathbf{u}_{\sigma_{-k}}, \mathbf{u}_{\sigma_0}, \mathbf{u}_{\sigma_k}\right) \right\}, \tag{1.16}$$

more than one pitch period on both sides of a concatenation point is taken into account, which allows measuring a signal dynamics throughout the whole concatenation area. Thus, (1.16) could be expected to have better performance in prediction of audible discontinuities than (1.15) [Bel04].

Evaluation

The performance of the proposed method was evaluated via a simple preference test in which the commonly used concatenation cost function based on the calculation of the Euclidean distance between MFCC vectors (39–dimensional MFCC vectors including delta and delta–delta features) played the role of a reference.

In contrast to other related studies, the test stimuli was not limited to concatenations in the middle of vowels (/A/,/u/ in SAMPA notation in this case) but also included concatenations in the middle of steady spectrum consonants (/m/,/n/), concatenations in the middle of diphthongs (/OI/, /aU/), and concatenations in the middle of varying spectrum consonants (/l/, /r/). The utterances included in the test stimuli were energy-normalized, and also the effect of pitch differences at concatenation points was limited by a simple pitch modification algorithm.

The best and the worst artificial concatenations were chosen by both of the methods and served as material used in the listening test. Seven listeners (5 generally conversant in speech processing, 2 with a more advanced background in psycho-acoustics and phonetics) were chosen to participate in the listening test. Their task was to judge whether the transition at the diphone boundaries was decisively smoother, about the same, or decisively more discontinuous when comparing the utterances presented pairwise in a randomized order.

The results of the preference test suggest that the proposed method considerably outperforms MFCC based concatenation cost function in all cases [Bel04]. Nevertheless, the results obtained for the liquids were not as convincing as for the other sets because the listeners were inclined to prefer none of the utterances.

1.4.4 Concatenation Cost Based on Non-Linear Methods

Introduction

The idea of applying non-linear speech modelling methods for the design of a concatenation cost function was presented in [PS05] and [PS07]. The authors assumed that a concatenation of two non-contiguous speech segments result in a signal of non-stationary characteristics which are difficult to accurately model by linear models. Instead, they suggest using features derived from a non-linear harmonic speech model and from a non-linear speech analysis algorithm. In the following paragraphs these methods are briefly described.

A Non-Linear Harmonic Model

The non-linear harmonic representation of speech models a speech signal as a sum, $h[n]$, of harmonically related sinusoids:

$$h\left[n\right] = \sum_{k=-L(n_i)}^{L(n_i)} A_k[n] \exp^{j2\pi k f_0(n_i)(n-n_i)}, \tag{1.17}$$

where $L\left(n_i\right)$ denotes the number of harmonics at $n = n_i$, $f_0\left(n_i\right)$ stands for the fundamental frequency at $n = n_i$, and $A_k\left[n\right]$ is defined as follows

$$A_k\left[n\right] = a_k\left(n_i\right) + \left(n - n_i\right) b_k\left(n_i\right), \tag{1.18}$$

$a_k(n_i)$ and $b_k(n_i)$ are complex numbers which denote the amplitude of the k-th harmonic and its first derivative (slope), respectively.

The unknown values of $a_k(n_i)$ and $b_k(n_i)$ are estimated by minimizing a weighted time-domain least-squares criterion:

$$e = \sum_{n=n_i-T_0}^{n=n_i+T_0} w^2[n]\,(s[n]-h[n])^2, \tag{1.19}$$

where $s[n]$ denotes the original speech signal, $w[n]$ a weighting window, and T_0 the local fundamental period $(f_s/f_0(n_i))$ in samples.

AM&FM Decomposition

Based on the Teager-Kaiser energy operator [Kai90]:

$$\Psi\{x[n]\} = x^2[n] - x[n-1]\,x[n+1] \tag{1.20}$$

Maragos et al. [MKQ92] have developed the Discrete Energy Separation Algorithm (DESA) for decomposing an AM-FM modulated signal into its components, i.e. amplitude modulated (AM) and frequency modulated (FM) components. The version of DESA used in [PS05] can be described as follows:

$$G[n] = 1 - \frac{\Psi\{y[n]\} + \Psi\{y[n+1]\}}{4\Psi\{x[n]\}} \tag{1.21}$$

$$\Omega[n] \approx \arccos(G[n]) \tag{1.22}$$

$$|a[n]| \approx \sqrt{\frac{\Psi\{x[n]\}}{1-G^2[n]}} \tag{1.23}$$

where $y[n] = x[n] - x[n-1]$, $\Omega[n]$ is the instantaneous frequency and $a[n]$ is the instantaneous amplitude. This algorithm can be applied for a signal decomposition around resonances found by bandpass filtering with a Gabor filter of impulse response defined by:

$$h_G\left[n\right] = \exp\left(-b^2 n^2\right) cos\left(\Omega_c n\right), \tag{1.24}$$

where b and Ω_c control the bandwidth and the central frequency of the filter, respectively.

Concatenation Cost Design

Pantazis and Stylianou [PS05] applied both of the above mentioned non–linear techniques in their work.

They obtained features estimated by the harmonic model using a two pitch period long analysis window from the left and the right side of concatenation points while reducing their number by taking only the lower 4 kHz of the signal spectrum into account. Since the features obtained by the harmonic model are complex numbers, they used the absolute of the complex difference to measure their similarity.

To obtain features based on the AM&FM decomposition, they set the length of the analysis window to approximately 20 ms, and applied a filter bank of twenty Gabor filters. The value of b was set to 250, and the central frequencies were uniformly distributed between 250 Hz and 5 kHz. A set of features from both sides of a concatenation point was estimated, and a sum of absolute differences was used as a similarity measure.

Evaluation

For the evaluation of the non-linear concatenation cost functions, the same speech material was used as in [KV01], i.e. five Dutch vowels constructed by concatenating diphones C_iV and VC_j excised from non-sense words. The listeners were presented with isolated vowels (the surrounding consonants were removed) with duration normalized to 200 ms, and their task was to make a binary decision regarding the presence of a discontinuity.

The performance of the concatenation cost function was then evaluated using discontinuity detection rates. The false alarm level was set to 5 %. The best performance was obtained for the harmonic model parameters combined with the AM&FM components, for which the detection rate was 56.35 %. Note that the performance of the traditional concatenation cost functions was considerably worse, reaching only about 40 % [PS07]. Still, the performance of the concatenation cost function based on the non-linear speech modelling techniques was found to be unsatisfactory.

1.4.5 Wavelet-based Concatenation Cost

Applying wavelets for the design of the concatenation cost functions is not a well documented field in literature. This technique was mentioned in [KOS06a] where some promising results have been presented. Using the wavelet analysis, the time-frequency resolution limitations of the Fourier transform may be overcome, which could be a beneficial feature for the spectral distance measures.

Many different wavelet functions exist, and a number of them were found to be giving similar results for the task of detecting the concatenation discontinuities [KOS06a]. Compared to the traditional discontinuity measures, the wavelet-based concatenation costs were giving slightly better results. The traditional concatenation cost functions considered in that work were the MFCCs, Log Power Spectra, cepstral coefficients and LSFs combined with the absolute distance, the Euclidean distance, the Cos distance, and the Symmetric Kullback-Leibler distance.

1.4.6 Phonetically-based Concatenation Cost

One of the alternative ways of avoiding audible discontinuities at concatenation points in synthetic speech is to take phonetic aspects into consideration. Different approaches are summarized in the following sections.

Humans' Discontinuity Detection Rates

In [Syr01], a study on human detection of discontinuities with respect to phonetic contexts of vowels was presented. It has been shown that the discontinuity detection rates are significantly higher when concatenating diphthongs, compared to short and long monophthong vowels. This comes as no surprise since the diphthongs contain two successive targets and there is a transition between them somewhere in the middle, i.e. at the location of a concatenation point. It is likely a spectral change at the concatenation area which accounts for the higher discontinuity detection rates in this case.

In the same work, there has also been shown that a post-vocalic context has considerably larger effect on discontinuity detection rates than a pre-vocalic context. This could suggest that anticipatory (right-to-left) coarticulation is stronger than retentive (left-to-right) coarticulation (see [KM77] for a discussion on this topic). Also a nature of a post-vocalic consonant has been found to play an important role. In Chapter 4 of this thesis, a detailed analysis of phonetic dissimilarity effects on quality of concatenations in vowels is presented. Our analysis disconfirms the observations made by [Syr01] in some parts.

In another study [SC05], the authors have shown that the mid-consonant discontinuity detection rates are approximately 20 % lower than for the mid-vowel joins reaching in average 46.6 %. They have however been found to vary to a large extent across different consonant groups as shown in Tab. 1.1.

All these findings are unquestionably an interesting factor, albeit not directly addressing the issue of measuring the concatenation discontinuities, to be taken into account during the design of a concatenation cost function.

One could for example consider introducing a phoneme-based concatenation penalization scheme to lower the number of concatenations at high risk locations or moving the concatenation points from the mid-phoneme area[5].

[5] A few boundary training techniques will be presented in Sec. 1.7.

Broad phonetic class	Joins detected [%]
Liquid sonorant	54.7
Glide sonorant	54.7
Aspirated glottal fricative	54.3
Nasal sonorant	47.4
Unvoiced weak fricative	35.3
Voiced weak fricative	23.3
Voiced strong fricative	22.2
Unvoiced strong fricative	13.6

Tab. 1.1: Concatenation detection observed in mid-consonant joins (after [SC05])

Phonetic Features as Discontinuity Predictors

Since no reliable acoustic measure has been found and widely accepted yet, some speech synthesizers supplement the concatenation cost function with phonetically based rules [Lee01].

The issue researched in [KT02] was to what extent can phonetic features be useful for the prediction of audible discontinuities. The authors trained statistical models using various sets of predictors and compared their discontinuity prediction abilities. First, they trained models using solely acoustic predictors, then solely phonetically-based predictors, and finally a combination of both.

The tested acoustic predictors were the MFCCs combined with the Euclidean distance, and the following list of the phonetic features played the role of phonetic predictors:

- manner of articulation of the succeeding consonant (*plosive, affricate, fricative, flapped, nasal, approximant*)
- place of articulation (*bilabial, dental, alveolar, alveolo–palatal, palatal, velar, uvular, glottal*)
- palatalization (*yes, no*)
- voicedness (*voiced/unvoiced*)
- equality of extraction/use environment
- tongue position of vowels (*front, mid, back*)

- tongue height (*high, low*)
- identity of V (*vowel, syllabic nasal*)

The performance of the trained models was evaluated using a set of 168 Japanese two-syllabic non-sense words in the form of VCV. From each word the initial vowel (all Japanese vowels and the syllabic nasal, more precisely) was extracted and concatenated with all the final CV parts available. The resulting set of words was then used as test stimuli for listening tests. The task of listeners was to rate the naturalness of the words on a 7-point scale.

The presented results indicate that the phonetic features significantly outperform the acoustic ones. It must, however, be noted at this point that the results obtained for acoustic measures were worse than presented in a similar study by [DFC98]. This partial disaccordance may be explained by different settings of conditions of the experiments, especially the difference between listening test stimuli. Surprisingly, no considerable improvement in the performance of the models was gained for the combination of both the acoustic and the phonetic predictors, compared to using solely the phonetic predictors.

1.5 Review of Presented Results

This section serves to summarize most of the results (see Tab. 1.2), which have been presented to date in the field of designing concatenation cost functions. The listed results are difficult to directly compare because the conditions of the presented experiments varied to a large extent. Nevertheless, it is obvious that some of the studies presented conflicting results. For each study, we report the speech material used for the evaluation of tested concatenation cost functions, the best concatenation cost function, the worst concatenation cost function and other concatenation cost functions within the scope of the particular work.

For example, the MFCCs combined with the Euclidean distance performed well in many studies, but in [KV01] their performance was found to be barely above chance level. On the other hand, [KV01] reported the Kullback-Leibler

Work	Speech material	Best	Worst	Others
Bellegarda [Bel04] [Bel06a]	/mAn/, /sun/, /Anu/, /umA/, /Alu/, /Aru/ /maUs/, /lOIn/	LSM–Based	MFCC + Eucl. dist.	/
Bjørkan et al. [BSF05]	sentences mid–vowel /A:/,/E:/	MFCC+ Eucl. dist. Cepst. dist.	LPC + KL dist.	pitch synch. crosscorr.
Kawai, Tsuzaki [KT02]	VCV join V–CV isolated	Phon. based cost	MFCC based cost	/
Kirkpatrick et al. [KOS06a]	the same as [SS01]	MFCC + Eucl.	LPC Power + Eucl.	all trad. costs
Kirkpatrick et al. [KOS06a]	the same as [SS01]	Wavelet based cost	/	all trad. costs
Kirkpatrick et al. [KOSE07a]	the same as [SS01]	MFCC + Eucl.	LSF + Eucl.	/
Klabbers, Veldhuis [KV01]	isolated Dutch vowels	LPC Power + KL dist.	MFCC + Eucl.	formants, Part. Loud. Itak. dist.
Pantazis, Stylianou [PS07]	the same as [KV01]	AM&FM + Fisher Lin.	LSF + Mahal.	all trad. costs
Pantazis, Stylianou [PS05]	The same as [KV01]	Non–lin. based cost	pow.–norm. spec. + KL dist.	/
Stylianou, Syrdal [SS01]	Modified Rhyme Test [HWHK63]	FFT pow. + KL dist. MFCC + Eucl.	LSF + Eucl.	all trad. costs
Vepa [Vep04]	sentences /eI/,/oU/,/aI/ /aU/,/oI/	LSF + Mahal.	Kalman based cost	MCA + Mahal.

Tab. 1.2: Summary of the results presented in the field of the concatenation costs design.

distance applied on the LPC power spectra to be the best predictor of the concatenation discontinuities, while in [KOS06a] or [BSF05], the same measure was found to be the worst. The study [SS01] found both of these measures to be the best predictors.

Another measure, for which conflicting results were presented, is the concatenation cost function based on the LSF parametrization. Vepa [Vep04] found the Mahalanobis distance between LSFs to perform the best, whereas in [PS07], the same measure was found to be the worst of all.

It is also worth mentioning at this point that the "alternative" methods (i.e. the non-linear techniques, the wavelets or the LSM-based approach) seem to be slightly outperforming the conventional FFT power-based approaches. It is however necessary to keep in mind that the performance of these techniques have not been confirmed upon their publication by any further studies.

In Sec. 2.6, a comparison between the LSM-based approach described in Sec.1.4.3 and the concatenation cost function based on the MFCCs combined with the Euclidean distance can be found.

1.6 Issues in Concatenation Cost Design

1.6.1 Role of Phase Mismatch

Introduction

Most of the discontinuity measures proposed to date are based on different parameterizations of magnitude spectra obtained by the Fourier transform. The phase information is typically discarded, which may cause the inability of the traditional discontinuity measures to correlate well with human perception. This particular concern has been mentioned in [Bel06a], and it is supported by the results presented in [KOS06a]. Also in this thesis in Sec.4.2, we confirm that phase mismatches can play an important role in some contexts.

Unfortunately, it has not been shown whether or not the phase mismatches tend to co-occur with the magnitude spectra discontinuities. Our analysis however suggests that this does not always need to be the case.

Group Delay Functions

One of the possible ways of inspecting the phase spectrum is to use a group delay function, $\tau(\omega)$, which is given by [YS95]:

$$\tau(\omega) = -\frac{\mathrm{d}}{\mathrm{d}\omega}\left\{\arg\left[X(j\omega)\right]\right\} \tag{1.25}$$

The group delay function can be computed directly from a signal using the following formula [MG03]:

$$\tau(\omega) = -\left(\frac{\mathrm{d}\left(\log\left(X(\omega)\right)\right)}{\mathrm{d}\omega}\right)_I = \frac{X_R(\omega)\,Y_R(\omega) + Y_I(\omega)\,X_I(\omega)}{\left|X(\omega)\right|^2}, \tag{1.26}$$

where the subscripts R and I denote the real and the imaginary parts, respectively. $X(\omega)$ and $Y(\omega)$ are the Fourier transforms of a discrete-time real signal $x(n)$ and $nx(n)$.

Estimating group delay functions is sensitive to a noise present in the signal, a window shape and also its length [BDDD04]. In [KOS06a], the best performance of phase-based discontinuity measures was obtained using pitch synchronous analysis with the Blackman window of the length of one pitch period. It has been found that phase-based discontinuity measures correlate with human perception, but their performance is slightly inferior to the traditional static magnitude spectrum measures [KOS06a].

1.6.2 Role of Spectral Dynamics

Introduction

The spectral dynamics plays an important role in human perception [Fur86], and can likely be a source of audible concatenation artifacts.

Common concatenation cost functions are however based on estimating the audible discontinuities by using two frames of signal neighbouring to the concatenation points (as shown in Fig. 1.2). This approach is obviously limited to measuring static differences at the concatenation points rather than capturing spectral dynamics of the whole concatenation area.

There are a few works dealing with this particular issue, [KOSE07a], [Bel04] or [VK06] to name some. In Chapter 3, the role of $F0$ dynamics at the concatenation points is investigated.

Delta Coefficients

In [VK06], delta coefficients, which are an estimation of time derivatives of static features, were used to represent a spectral change together with a static spectral parametrization. It has shown that the inclusion of the delta coefficients is only of a minor benefit to the performance of the discontinuity measures while the memory requirements are increased considerably.

Nevertheless, similarly to all aspects of the concatenation cost design, there is a work which reports contrary results. In [Don01], the inclusion of the delta and the delta-delta coefficients has been found to be beneficial for the performance of the traditional concatenation cost functions.

Average Multi-Frame Distances

An alternative to using the delta coefficients is a calculation of multi-frame distances [Vep04]. Fig. 1.6 illustrates how a three-frame spectral distance can be obtained. Vepa presented in his work the results of experiments using three-, five-, seven- and nine-frame spectral distances using MCA, MFCC and LSF

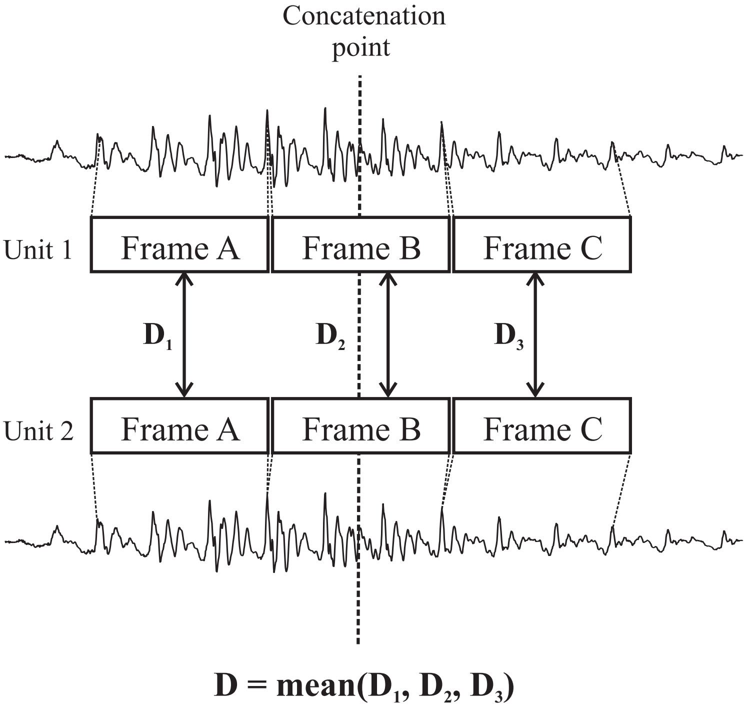


Fig. 1.6: Three–frame spectral distance

The three-frame distance between two units is calculated as an average of distances between corresponding frames of each unit.

coefficients combined with the Euclidean, the absolute, the Mahalanobis and the Kullback-Leibler distances. It has been found that single frame distances outperform those of multi-frame in most cases, and if a small increase in correlations is found, it is overweighted by a considerable increase in memory requirements.

Spectral Trajectory Modelling

The idea of spectral trajectory modelling is based on using polynomial coefficients which fit the best the static parameter trajectories [KOSE07b], [KOSE07a]. The static parameters can be calculated at each pitch mark[6] in the concatenation area, and modelled by a polynomial function. The spectral dynamic parameters are then obtained as derivatives of the polynomial functions at the concatenation point.

Kirkpatrick et al. [KOSE07b], [KOSE07a] have experimented with the modelling of trajectories of various spectral parameters, including MFCCs and LSFs or formant frequencies, and obtained similar results for all of them. In addition to the first derivatives, they also calculated the second derivatives. As a baseline, they used the delta coefficients. All these dynamic features were also compared with the traditional static parameters.

They have found the spectral trajectory estimates to correlate with human perception. Nevertheless, the correlations of static parameters have been found to be higher. Surprisingly, the performance of the delta coefficients was just above the level of pure chance. Similarly to [WM98] and [VK06], combining static and dynamic parameters gave only minor improvements. This has been explained by the high correlation of these parameters, suggesting that the static and the dynamic mismatches tend to co-occur [KOSE07a].

1.6.3 Windowing for Feature Extraction

It has been found [KOS06a], [KOS06b] that the setting of a window length for the extraction of features also plays an important role, and can even be more critical for the quality of the discontinuity measures than the selection of a spectral parametrization itself, which was the main focus of many of the related studies. The obtained results suggest that correlations with human perception vary significantly more when altering the window length than when using different spectral parameterizations.

The optimal strategy for the extraction of the traditional static features has been found to be the pitch synchronous windowing with the window length

[6] Pitch mark can be defined as the location of a speech signal amplitude extreme (peak or valley) that corresponds to a moment of a glottal closure [LMT11].

equal to one pitch period. For the non-pitch synchronous extraction, the window length close to an estimated average pitch period of the speaker is the best option [KOS06a].

In contrast to the static features, it seems to be more appropriate to use longer windows for measuring the dynamic features described in Sec. 1.6.2. A possible explanation is that smoother trajectories of parameters are obtained using longer windows, which allows better capturing of the signal dynamics [KOSE07b].

1.7 Optimal Boundary Training

1.7.1 Introduction

It has already been mentioned in Sec.1.4.6 that some techniques for lowering the probability of audible concatenation artifacts exist. In addition to phonetically based methods, which also include techniques for enriching unit inventories with context sensitive units [KV01] and/or using longer units to lower a number of concatenation points [Sag88], [CBd09], smoothing techniques can also be used [Vep04], [LK02]. These techniques however introduce an inevitable risk of a loss of naturalness of synthetic speech [CH98]. Recently, methods for combining model and signal based synthesis techniques have also been introduced [PB08].

Boundary training (also referred to as "optimal coupling") techniques represent another option. These techniques are based on searching for locations within phonemes where concatenations are supposed to be more feasible than concatenating at midpoints. The limitation of the boundary training methods is given by the impact on phoneme durations, they may have.

It is worth mentioning a very important fact at this point. The signal based techniques rely on knowledge, obtained by objective measurements, of what is perceptually important. This is basically the same issue as for the concatenation cost design.

In the following sections, two boundary training methods will briefly be described.

1.7.2 Simple Frame Mismatch Method

To our knowledge, the first boundary training method was introduced in [CI97]. It is based on an idea of using a simple frame mismatch measure to find the most appropriate point for concatenating. The method can be summarized as follows.

The MFCCs are extracted from both of the units being concatenated using a 25.6 ms long analysis window and 5 ms frame shift. Then, the Euclidean distance is applied to measure pairwise similarities between extracted coefficients. The optimal boundary position is then found as a point that corresponds to the minimum distance between parameterizations, i.e. the resulting synthetic unit is composed of the left part of the first unit up to the frame f_i, and the right part of the other unit from a frame $\bar{f}_j$, where frames f_i and $\bar{f}_j$ are those giving the minimum distance. See Fig. 1.7 for an illustration of the method.

The results presented in [CI97] suggest that this technique is capable of finding better positions for the cut points, which results in an improvement of the quality of synthetic speech. It has unfortunately also been found to be introducing considerable changes of durations of synthesized units. Thus, the algorithm has to be tuned to find the best trade-off between the concatenation quality and the duration modifications by setting search region limits. Alternatively, the durations of the synthesized units have to be modified by signal modification techniques (e.g. [VR93]).

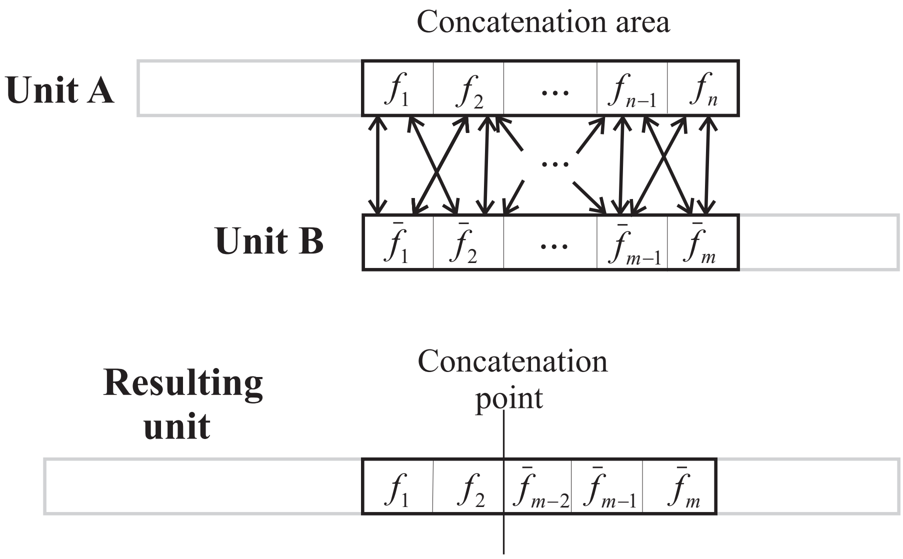

Concatenation area
Unit A
f_1 f_2 ... f_{n-1} f_n
...
Unit B
$\bar{f}_1$ $\bar{f}_2$ ... $\bar{f}_{m-1}$ $\bar{f}_m$
Resulting unit
Concatenation point
f_1 f_2 $\bar{f}_{m-2}$ $\bar{f}_{m-1}$ $\bar{f}_m$

Fig. 1.7: Optimal coupling method based on the simple frame mismatch

When concatenating two units, the optimal concatenation point is found by measuring a difference between parameterizations taken from both units. The units are concatenated at a point which corresponds to the minimum distance. If no search region limits are set, it may result in significant changes in durations of synthesized units.

1.7.3 LSM-based Boundary Training

LSM-based boundary training is a technique built upon the feature extraction framework based on the latent semantic mapping (LSM) paradigm [Bel05], [Bel04]. This feature extraction framework has already been described in Sec. 1.4.3.

The idea of boundary training consists in applying an iterative procedure depicted in Fig. 1.8. It adjusts individual boundaries using a global concatenation criterion (1.16) in each turn until convergence [Bel06a]. For a proof of convergence of the proposed iterative procedure, refer to [Bel06b].

The boundary training can be done as an off-line procedure. At synthesis runtime, the found boundaries can then be considered as fixed, and only the concatenation costs need to be calculated.

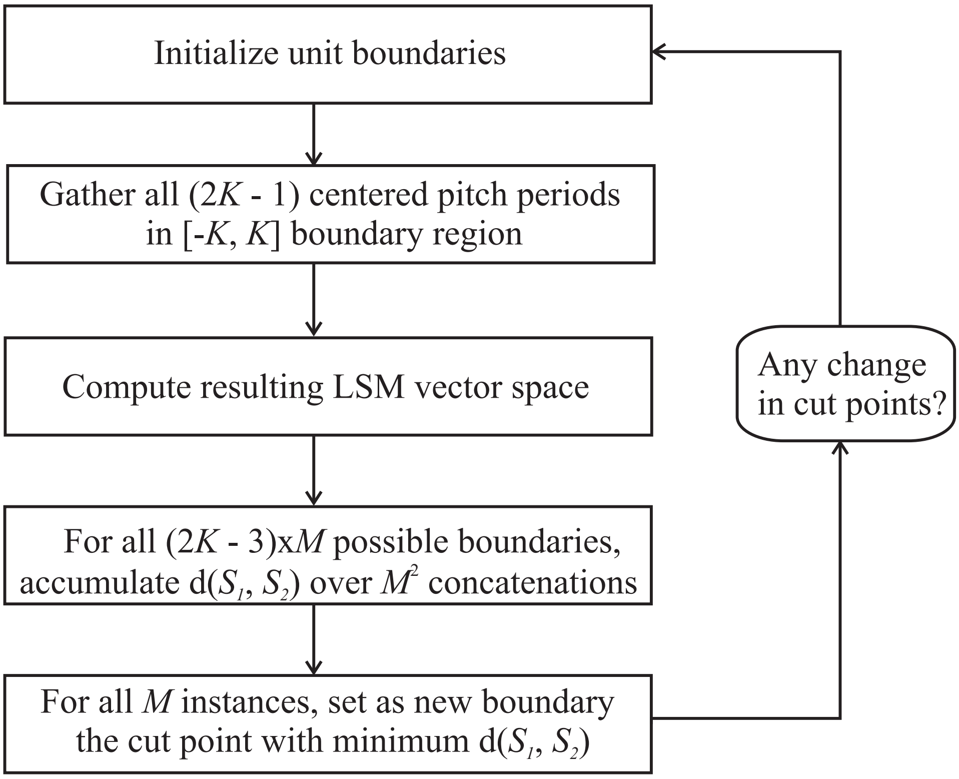

Initialize unit boundaries
Gather all (2K - 1) centered pitch periods in [-K, K] boundary region
Compute resulting LSM vector space
For all (2K - 3)xM possible boundaries, accumulate d(S1, S2) over M2 concatenations
For all M instances, set as new boundary the cut point with minimum d(S1, S2)
Any change in cut points?

Fig. 1.8: LSM–based iterative unit boundary training (after [Bel06b], [Bel06c]).

A flowchart of an iterative procedure for refining unit boundaries. M stands for a number of units which contain a boundary of a given phoneme, the constant K determines how many pitch periods are considered as a concatenation area, and the value $(2K-3)$ specifies an allowed range for a cut point shift in one iteration. For more details refer to Sec. 1.4.3.

1.8 Evaluation of Concatenation Cost Functions

1.8.1 Introduction

The task of designing a concatenation cost function is in fact searching for a measure that highly correlates with human perception of concatenation artifacts. It is the complexity of human perception that makes the evaluation of the concatenation cost functions a difficult problem as it brings to the task a lot of subjectivity and psychoacoustic issues, which represent separate research areas themselves.

Generally speaking, there are two typical ways of the evaluation of the concatenation cost functions. First, one can have a set of concatenation cost functions, synthesize the same sentences using each of them separately, and then ask listeners to choose the best version, or to compare the synthesized and natural versions of the same sentences. This approach can especially be useful for a fine-tuning of weights or any parameters of the concatenation cost functions as the test stimuli play a role of calibration sentences at the same time.

The other option is to simply concatenate units, let listeners assess the quality of concatenation points, and then calculate correlations or detection rates of the discontinuity measures and the listeners.

1.8.2 Listening Tests Stimuli Considerations

Speech Material

It is obvious that the choice of a speech material used for the evaluation of the concatenation cost functions depends on a target phoneme group of the concatenation cost design.

Majority of the concatenation cost functions proposed so far have been tested on concatenations in vowels ([KV98], [BSF05] or [SS01] to name but a few). Vowels are the primary focus of our work too as they have been found to contain the largest number of audible concatenation artifacts [SC05].

In [Vep04], diphthongs were used, which was motivated by the assumption that if any spectral discontinuity measure gives good results on diphthongs, it likely also has a good performance on other phonemes.

Only in a few works, the spectral measures have been evaluated on other sounds. In [Bel06a], for instance, the steady spectrum consonants /m/, /n/ and the varying spectrum consonants /l/, /r/ were used in addition to vowels and diphthongs.

Length of Stimuli

In general, the length of individual stimuli depends on an aspect of the quality of synthetic speech a test designer is evaluating.

Shorter units (e.g. isolated phonemes) are often used for the evaluation of quality of concatenations. This is a natural choice as these units can be synthesized containing a single concatenation point, which is important for mitigating possible effects of surrounding concatenations on listeners' ratings of a concatenation point of interest. Also, impact of other disturbing elements (e.g. prosodic flaws) is limited.

Nevertheless, the disadvantage of this approach is that the stimuli are probably too short to allow listeners to perceive discontinuities and judge them consistently [KV01]. Even when using two-phoneme long stimuli with the concatenation point in the middle, the inter-rater correlation can be quite low (0.34 reported in [Don01]).

In our work, a "half sentence" method described in detail in Sec. 2.2 has been proposed.

Smoothing

Common practice in the evaluation of the concatenation cost functions is smoothing concatenation points with respect to $F0$ and energy differences [Bel04], [Lee01], [SC05]. The inevitable risk of smoothing is however a possible introduction of smoothing artifacts. For this reason, no smoothing techniques were applied in our work.

Instead, we measured the $F0$ and energy differences at concatenation points and took the differences into account when creating the initial listening test stimuli (see Sec. 2.6.2). In later stage, classification models trained on manually labeled data were used to factor out the $F0$ concatenation discontinuities. This was the case for the phonetic context analysis described in Chapter 4.

1.8.3 Reliability Analysis

Generally speaking, listening tests are still the only way to reliably assess quality of synthetic speech. Nevertheless, the key factor which is not completely under control are listeners themselves.

Reliability analysis of data collected as part of an evaluation of concatenation cost functions should ideally cover the *inter-rater reliability*, which aims at analyzing how much agreement exists among listeners rating the same test stimuli; the *internal-consistency reliability*, which is concerned with consistency of scores obtained for test items that represent the same construct; and also the *inter-method reliability*, which measures agreement between scores obtained for the same constructs by different methods, and which in fact represents the quality of a concatenation cost function under evaluation.

The issue of the inter-rater reliability is in works dealing with the concatenation cost functions often addressed either by an inter-rater correlation analysis [Don01] or by applying the general theory of signal detectability[7] [SC05].

The internal-consistency reliability can be estimated using natural [Ben05] and revision sentences as part of the listening test stimuli. Including natural sentences can also to some extend support the identification of less reliable listeners done by means of the inter-rater reliability analysis (see Sec. 2.6.4).

The most common methods of evaluating the performance of the concatenation cost functions (the inter-method reliability) include *correlation analysis* [Don01] when the correlation is calculated between scores given by a concatenation cost function under evaluation and Mean Opinion Scores (MOS) obtained by listening tests; *ROC analysis*[8], by which the results can be demonstrated graphically [BSF05], [KV98] or as a numeric value in a form of an area under the ROC curve (AUC) [KOSE07b], [KOSE07a]; and a *detection rate*

[7] The signal detection theory deals with quantifying the ability of discerning between signal and noise [Swe64].

[8] A receiver operating characteristics (ROC) analysis is a technique for the evaluation of performance of classifiers [Faw06].

calculation [PS07], [SS01], which is normally done using a fixed false alarm rate value. Last but not least, preference tests can also be used to make comparisons between different concatenation cost functions.

1.9 Case Studies

The goal of this section is to present a few examples of concatenations in order to demonstrate what sorts of phenomena can possibly occur at the concatenation points. Obviously, the examples cannot show all possible observations one can make when analyzing the concatenation points. They rather illustrate the subject of the thesis and the content of the analyses presented in the following chapters.

In Fig. 1.9, phase mismatch at a concatenation point is shown. To some extend, a discontinuity can also be observed in the power spectrum but the $F0$ contours of the concatenated units are reasonably coherent.

In contrast, Fig. 1.10 shows an example of an audible concatenation, which is probably due to the difference in slopes of the $F0$ contours of the concatenated units. In particular, one of the contours is rather stable at 145 Hz whereas the other one has an increasing trend from 145 Hz to 165 Hz within the region of eight pitch periods shown in the figure. Note that the static difference in $F0$ at the concatenation point is less than 5 Hz. Very slight discontinuity can also be observed in the power spectrum but this one is hardly measurable using the Euclidean distance of the MFCC vectors.

Finally, in Fig. 1.11, we show an example of an easily measurable discontinuity in the power spectrum. At the same time, the trends of the $F0$ contours are also rather incoherent.

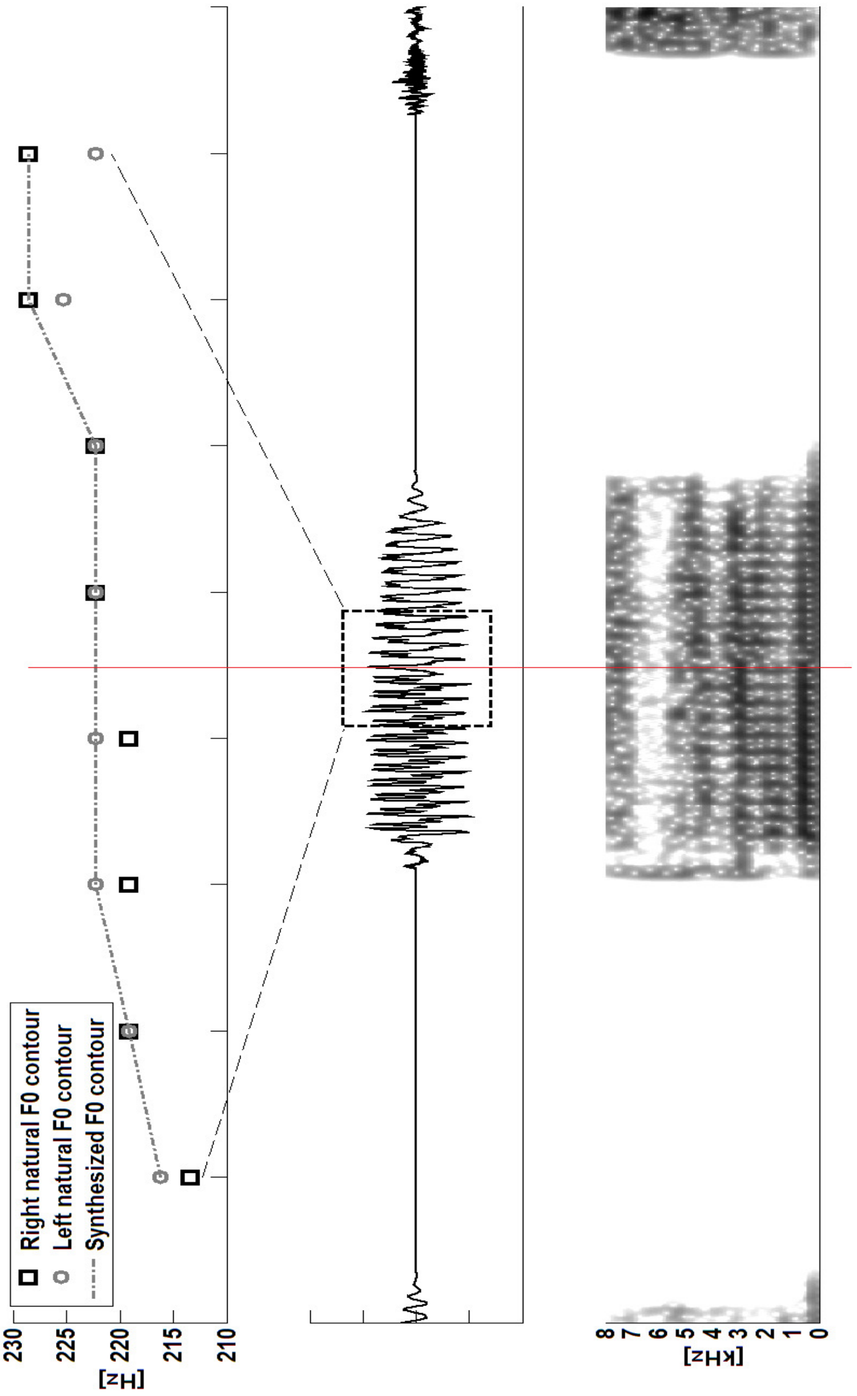


Fig. 1.9: Case study 1—Discontinuous concatenation in the middle of /e/, female voice

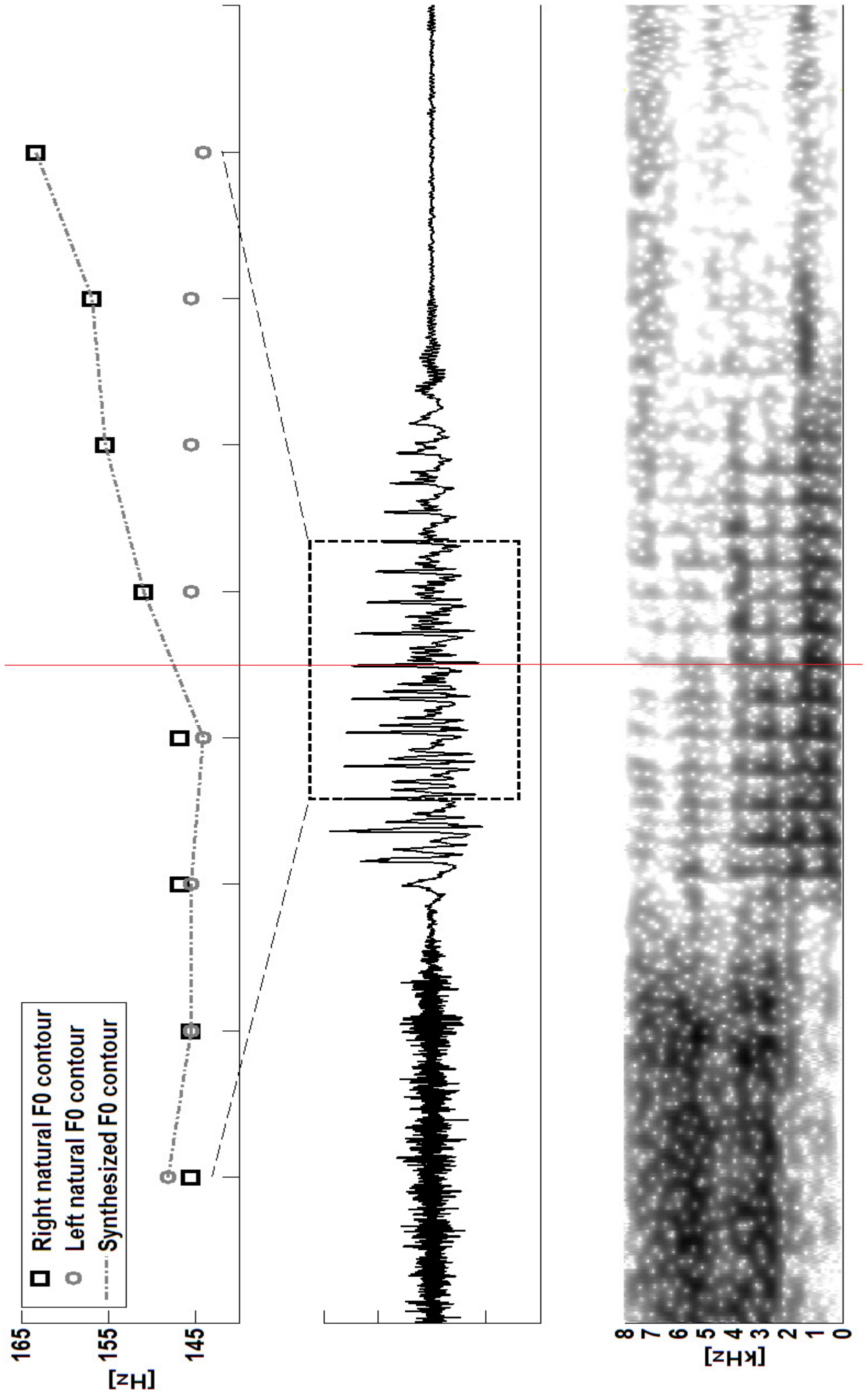


Fig. 1.10: Case study 2—Discontinuous concatenation in the middle of /a/, male voice

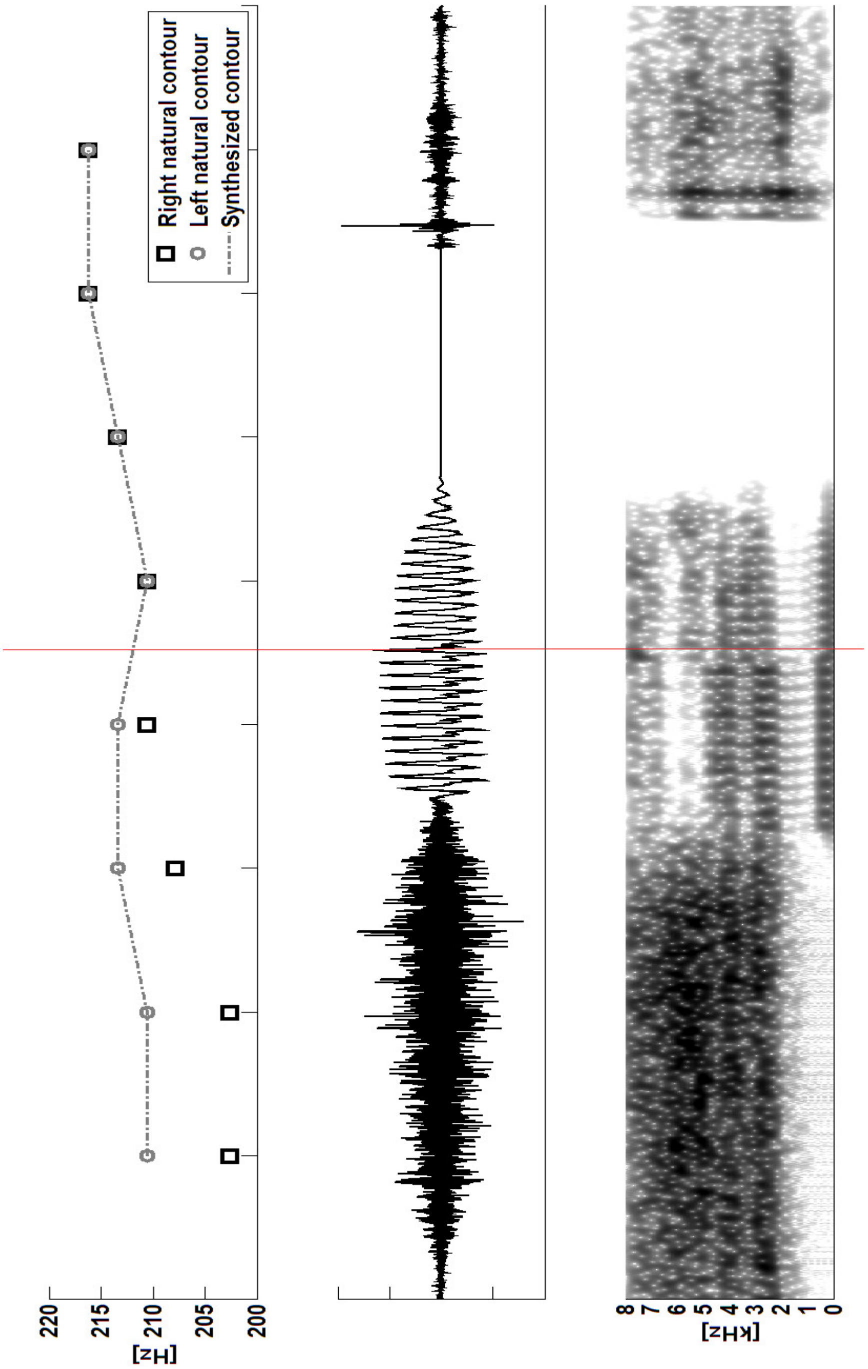


Fig. 1.11: Case study 3—Discontinuous concatenation in the middle of /i/, female voice

1.10 Scope and Structure of this Thesis

1.10.1 Objectives of this Thesis

It is obvious from the overview provided in the previous sections that a lot of work has already been done in the field of designing concatenation cost functions for the unit selection speech synthesis. Unfortunately, the results of this work are not satisfactory. This is mainly for two reasons. First, correlations between proposed objective measures and scores obtained in perceptual experiments are low. Second, there are disagreements between different studies.

One of the reasons for the inconsistencies found in the presented results can be a difference in evaluation data and evaluation approaches that were applied. The first objective of our work is therefore to propose a method for collecting reliable perceptual data. The goal is to obtain data that are consistently annotated by a large number of listeners and cover a wide range of possible concatenation discontinuities.

The second objective is to investigate the role of $F0$ in the perception of concatenation discontinuities. Our work is focused on comparing static and dynamic representations of $F0$. To our knowledge, such a comparison has never been done yet. The approach taken in previous studies was mainly to mitigate the role of $F0$ by applying different smoothing techniques and to concentrate on measuring spectral discontinuities. This is in our opinion another possible reason for the inconsistencies in the previous works. Apparently, if too much $F0$ smoothing is involved, audible artifacts that are difficult to measure in spectrum can be introduced and affect the results.

It has been shown that information about phonetic contexts of concatenated units can be leveraged to improve or to supplement concatenation cost functions [KT02], [Lee01], [Syr01]. Our aim is to confirm this finding and provide a description at signal level of what the context dependent coarticulation phenomena are that influence human perception.

Unit selection based TTS systems are known for their ability to generate almost natural synthetic speech. Haphazardly appearing quality jumps are however common to all such systems. Some of the audible artifacts can probably be avoided by introducing a better concatenation cost function as they appear at concatenation points. In the final part of our work, we concentrate on quantizing the potential quality gain that can be achieved by improving the quality of concatenations by selecting units more feasible for inaudible concatenations. We also consider other unit selection costs and their perceptual relevance.

The large amount of previously conducted studies also show that the problem of designing concatenation cost functions is a very complex one. Therefore, our primary focus is to measure concatenation quality in five short Czech vowels for two speakers—female and male.

Our work does not deal with smoothing methods nor optimal boundary training. For both of these topics, understanding of perceptually relevant phenomena related to the quality of concatenations is in our opinion a necessary prerequisite.

1.10.2 Thesis Structure

This thesis is divided into six chapters. The current chapter serves as an introduction to the problem and gives an overview of what the content of our work is. Chapter 2 describes a procedure that was used for collecting experimental data leveraged throughout different experiments described in this thesis. A method for synthesizing listening test stimuli is proposed and evaluated in that chapter. In Chapter 3, the role of $F0$ with respect to the quality of concatenations is investigated. The impact of consonantal contexts on the quality of mid-vowel concatenations is analyzed in Chapter 4. In Chapter 5, the work on the concatenation cost function design is put into a wider perspective of the unit selection method. An analytic algorithm that allows for measuring the perceptual relevance of unit selection costs and their sub-components is proposed. Finally, in Chapter 6, conclusions of this thesis are provided and an outline for future work is given.

1.10.3 Publications

Parts of this thesis were published as papers in conference proceedings or journals. The following is the list of the related publications:

Data Collection and Analysis

- M. Legát and J. Matoušek, "Analysis of Data Collected in Listening Tests for the Purpose of Evaluation of Concatenation Cost Functions," in *Text, Speech and Dialogue, proceedings of the 14th International Conference TSD 2011, Lecture Notes in Artificial Intelligence*, pp. 33–40, Springer, Berlin-Heidelberg, Germany, 2011.
- M. Legát and J. Matoušek, "Collection and Analysis of Data for Evaluation of Concatenation Cost Functions," in *Text, Speech and Dialogue, proceedings of the 13th International Conference TSD 2010, Lecture Notes in Artificial Intelligence*, pp. 345–352, Springer, Berlin-Heidelberg, Germany, 2010.
- M. Legát and J. Matoušek, "Design of the Test Stimuli for the Evaluation of Concatenation Cost," in *Text, Speech and Dialogue, proceedings of the 12th International Conference TSD 2009, Lecture Notes in Artificial Intelligence*, vol. 5729, pp. 339–346, Springer, Berlin-Heidelberg, Germany, 2009.

Design of Concatenation Cost Functions

- M. Legát, "Impact of Phonetic Context Mismatches on Quality of Vowel Concatenations," in *Proceedings of the 11th International Conference on Signal Processing*, pp. 523–526, Beijing, China, 2012.

- M. Legát and R. Skarnitzl, "The Role of Nasal Contexts on Quality of Vowel Concatenations," in *Text, Speech and Dialogue, proceedings of the 15th International Conference TSD 2012, Lecture Notes in Artificial Intelligence*, pp. 551–558, Springer, Berlin-Heidelberg, Germany, 2012.
- M. Legát and J. Matoušek, "Identifying Concatenation Discontinuities by Hierarchical Divisive Clustering of Pitch Contours," in *Text, Speech and Dialogue, proceedings of the 14th International Conference TSD 2011, Lecture Notes in Artificial Intelligence*, pp. 171–178, Springer, Berlin-Heidelberg, Germany, 2011.
- M. Legát and J. Matoušek, "Pitch Contours as Predictors of Audible Concatenation Artifacts," in *Proceedings of the World Congress on Engineering and Computer Science 2011*, pp. 525–529, San Francisco, USA, 2011.

Supportive Publications

- M. Legát, J. Matoušek, and D. Tihelka, "On the Detection of Pitch Marks Using a Robust Mutli-Phase Algorithm," *Speech Communication*, vol. 53, no. 4, pp. 552–566, April 2011.
- M. Legát, D. Tihelka, and J. Matoušek, "Pitch Marks at Peaks or Valleys?," in *Text, Speech and Dialogue, proceedings of the 10th International Conference TSD 2007, Lecture Notes in Artificial Intelligence*, pp. 502–507, Springer, Berlin-Heidelberg, Germany, 2007.
- M. Legát, J. Matoušek, and D. Tihelka, "A Robust Multi-Phase Pitch-Mark Detection Algorithm," in *Interspeech 2007*, vol. 1, pp. 1641–1644, Antwerp, Belgium, 2007.

Chapter 2

Collection of Experimental Data

2.1 Introduction

This chapter describes our data that were collected for the purposes of the evaluation and design of the concatenation cost functions. The data are used in different experiments described throughout the thesis.

We start with a proposal of a method that was used for creating synthesized data. The method is presented in Sec. 2.2 and its description includes the rationale for introducing new procedures for the perceptual data collection.

The method is verified and compared to one of the traditional approaches with respect to its feasibility for collecting reliable and consistent perceptual annotations. This can be found in Sec. 2.5. That section also addresses the problem of selecting the right test stimuli for listening tests aiming at collecting annotated data exploitable for our purposes.

In Sec. 2.3, the set up of the listening tests is presented. The description includes procedures that were used for conducting a listeners reliability analysis as well as an identification of annotated sentences that can be considered as containing "significant" observations[1].

[1] By "significant" we mean here sentences for which listeners find a good agreement regarding the concatenation quality.

Finally, in Sec. 2.6, the large scale listening tests are described and evaluated. This section also contains figures describing the collected data.

2.2 Half-Sentence Method

In Sec. 1.8.2, different aspects for consideration when preparing a listening test for the evaluation of concatenation cost functions were discussed. One of these aspects was the length of test stimuli. Since neither very short stimuli containing a single concatenation point nor longer stimuli with multiple concatenations are ideal, we decided to propose a new method for our experiments. We call this method *half-sentence method*.

As a preparation, a database of short Czech sentences containing three words each, e.g. `/kra:lofski: kat konal/` (SAMPA notation), was recorded. The middle words of these sentences are of our interest as they are in the form consonant-vowel-consonant (CVC), i.e. words containing a vowel surrounded by two consonants. These sentences were designed to cover all short Czech vowels (/a/, /e/, /i/, /o/ and /u/) in all possible consonantal contexts. The sentences were uttered by two speakers—female and male.

Then, each sentence was cut into halves, i.e. the cut points were in the middle of the mid-vowels of the central mono-syllabic words. All possible combinations of left and right halves were made as depicted in Fig. 2.1. When concatenating the halves, no smoothing technique was applied for the reasons discussed in Sec. 1.8.2. Only a simple overlap and add method was used using an overlapping region of 5 ms weighted by the Hanning window.

It was expected that this approach would provide us with data, which could be reliably rated by a larger group of not necessarily expert listeners as the stimuli was long enough and containing a single concatenation point at the same time. To make the task of the listeners even easier, it was also planned to present them with natural versions of the synthesized middle words prior to playing the whole sentences. In order to verify the feasibility of the proposed

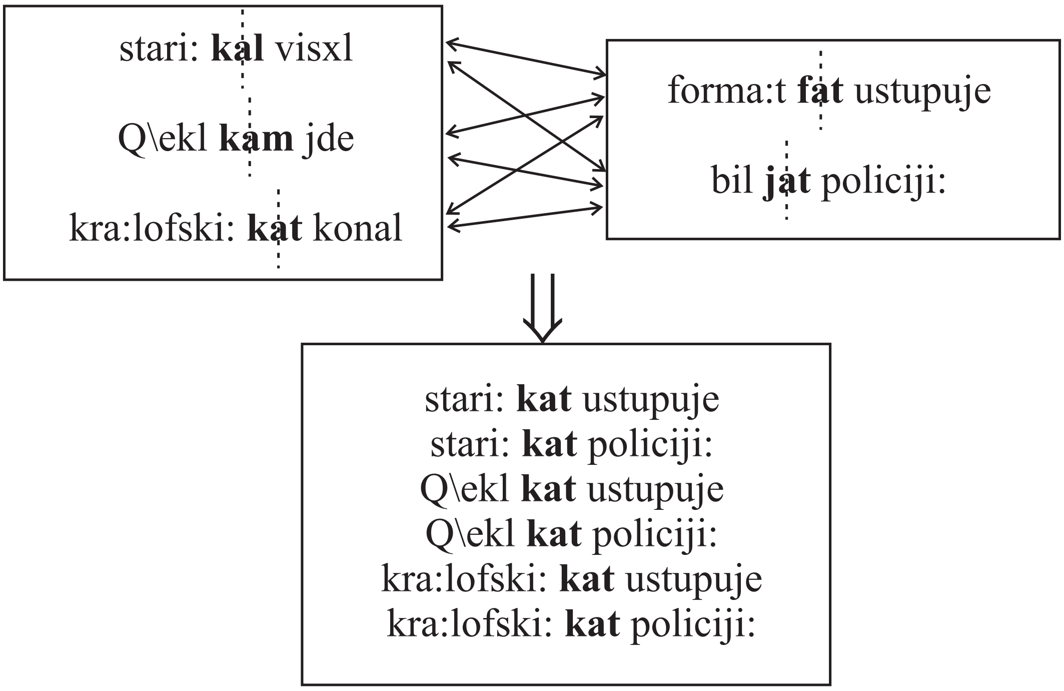


Fig. 2.1: Synthesis of sentences using the *half-sentence method.*

Given a reference word (`kat` in this case), test sentences are composed of left and right parts of the recorded sentences containing the corresponding parts of the reference word.

approach, we conducted a preliminary listening test, which will be described in Sec. 2.5.

2.3 Listening Tests Setup

2.3.1 Subjects

The listeners participating in our listening tests were in vast majority university students, all native speakers of Czech. Two of the participants were speech synthesis experts, a few others stated that they had some background in phonetics. Approximately half of the listeners were involved in listening tests for both speakers, some were also participating in the preliminary listening test. All subjects were paid upon completion of the tests.

2.3.2 Procedure

The listeners were allowed to play each audio sample as many times as they liked before providing their ratings. The task was to assess the concatenations in the middle vowel of the central word of each sentence using both a five-point scale:

- *no join at all*
- *unnatural but not disturbing*
- *slightly perceived join*
- *highly perceived join*
- *highly disturbing join*

and a binary scale:

- *perceived join*
- *not perceived join*

All listening tests were conducted using a web interface allowing the listeners to work from home. It was, however, stressed in the test instructions that the tests shall be done in a silent environment and using headphones. It is obvious that organizing the tests in a laboratory would provide us with more consistent testing conditions, but taking into account the number of listeners and the length of each listening test, especially of the large tests described in Sec. 2.6, it would have been unacceptably time-costly.

To gain more control over the listeners, not only were logs from our test server analyzed but also some control mechanisms were included into the tests themselves. These control mechanisms and the listeners' reliability analysis are described in the following section.

2.4 Listeners Reliability Analysis

2.4.1 Analysis of Testing Tool Logs

The test framework used for conducting the listening tests allows for checking actions each listener performs during conducting listening tests. This helps to identify listeners who spend extremely short time to complete a test compared to others, which suggests that they might have been providing their ratings not thoroughly enough. It is also possible to see whether or not the audio samples are played before being rated.

All ratings given by listeners who were found to be suspicious based on the analysis of the testing tool logs were discarded. They were only kept for experimental reasons in the further reliability analysis described in the following subsections in order to see if they would be detectable without having the possibility of investigating the test server logs[2].

2.4.2 Rating of Natural Sentences

As also found in [Ben05], inclusion of natural speech samples is a valuable resource for identifying malicious listeners. There were some natural sentences included in all our tests and used during the evaluation to estimate the ability of the listeners to identify natural speech. Those, who assessed a natural sentence as containing an audible join, were given one penalty point for each such decision. In addition, the five-point scale ratings were also checked, and each listener was given a score using a penalization scheme shown in Tab. 2.1. The listeners were then ranked according to their scores, and a small group of deviating listeners could be found in each test.

[2] This should serve as a verification of the appropriateness of the reliability analysis procedures.

Tab. 2.1: Penalization scheme based on the listeners' answers using the five-point scale. **Diff** stands for a difference from one, which was the expected rating of natural sentences.

Diff	**Penalty**
0	0
1	-0.001
2	-0.01
3	-0.1
4	-1

2.4.3 Internal-Consistency Reliability

In order to measure the internal consistency, revision sentences in the form of double appearances of randomly selected sentences were included into the test stimuli of each test. The method for scoring the listeners based on the revision sentences was the same as for the natural sentences. The only difference was that the reference rating was given by the rating of the firstly appearing sentence in each pair.

All inconsistent decisions on the binary scale, i.e. rating the first occurrence as continuous and the other as discontinuous or vice versa, were penalized by one penalty point. Any difference in scores given by the five-point scale was penalized according to Tab. 2.1.

It is worth noting at this point that some sentences were found to be ambiguous in their nature as they contained hardly perceivable joins. This was an explanation of lower scores all listeners obtained for the revision sentences in comparison to the scores based on the natural sentences.

2.4.4 Inter-Rater Reliability

Distribution of Ratings

The listeners were instructed to use the full range of levels when assessing the joins on the five-point scale. In addition, the large scale listening tests

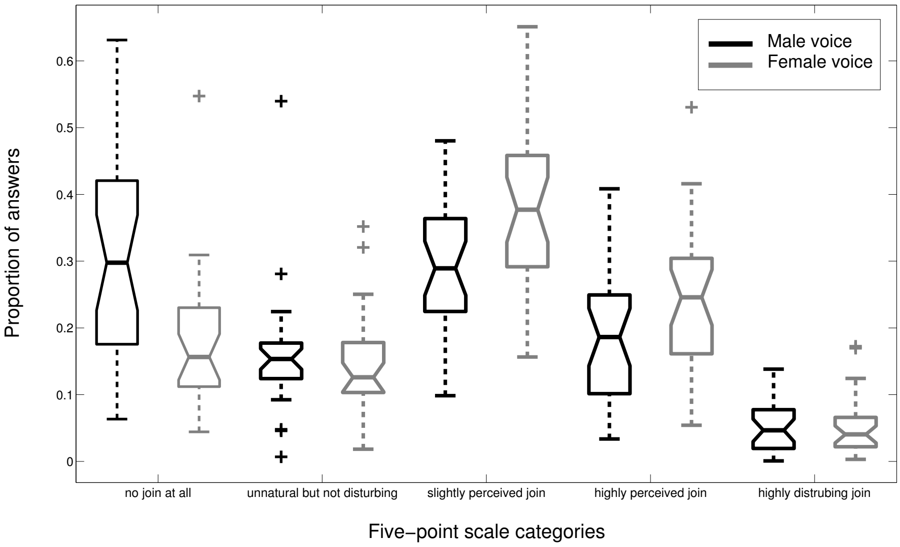


Fig. 2.2: Distribution of listeners' assessments on the five-point scale.

Each box-and-whisker plot shows ratios of all listeners' ratings falling within a given category. The outliers (shown as plus symbols) identify listeners who were inclined to rate more/less frequently using a given category. For instance, this particular plot identifies a listener who was rating for the female voice more than 50 % of the time using the category "no join at all", which was deviating from all other listeners.

contained a calibration set of sentences to train the listeners before starting the tests, and to re-calibrate whenever needed.

To verify that all listeners indeed went through the calibration phase and also that the calibration was effective, distributions of the five-point scale ratings were analyzed. An example of a result of this analysis is given in Fig. 2.2. The figure shows box-and-whisker plots of ratings given by the listeners in both large scale tests.

(Dis)agreement with Facts

In order to evaluate the inter-rater reliability with respect to the binary scale ratings, a set of sentences that were assessed by a majority of listeners in the same way using this scale, either as containing an audible join or being completely natural, was collected in each listening test. Henceforth, these sentences will be referred to as *facts*. The *facts* can formally be described as

$$sent_i \in facts \quad \Leftrightarrow \quad \frac{N_i^+}{N_i} \geq 0.8 \ \vee \ \frac{N_i^-}{N_i} \geq 0.8, \tag{2.1}$$

where $sent_i$ is the i-th sentence of the test stimuli, N_i^+ and N_i^- are the numbers of continuous (i.e. *not perceived join*) and discontinuous (i.e. *perceived join*) ratings given to the i-th sentence, respectively, and N_i is a total number of ratings given to the i-th sentence. The value 0.8 represents 80 % of ratings, and it was set ad hoc[3].

Having the collected *facts*, it is possible to iterate and to calculate a disagreement score for each listener using the following formula:

$$\mathit{DISAGR}_i = \frac{\mathit{NUM\_DIFF}_i}{\mathit{FACT\_COUNT}} \times 100\,[\%]\,, \tag{2.2}$$

where DISAGR_i is the disagreement score of the i-th listener, $\mathit{NUM\_DIFF}_i$ is the number of assessments of the i-th listener different from the *facts* collected by all listeners excluding the i-th listener, and the $\mathit{FACT\_COUNT}$ is the number of collected *facts*.

Starting with all listeners, a few iterations were sufficient to find those who deviated from the majority. These listeners can be seen in the course of the metric based on the measure (2.2). In each iteration, the worst listener was excluded, and a new set of *facts* was created. Fig. 2.3 shows a plot of this metric as it looked like for the large-scale listening tests.

The steps of one iteration of the disagreement analysis can be summarized as follows:

1. Exclude ratings of the first listener in the set.

[3] Obviously, a more sophisticated approach could have been taken but it was not the primary focus of this work.

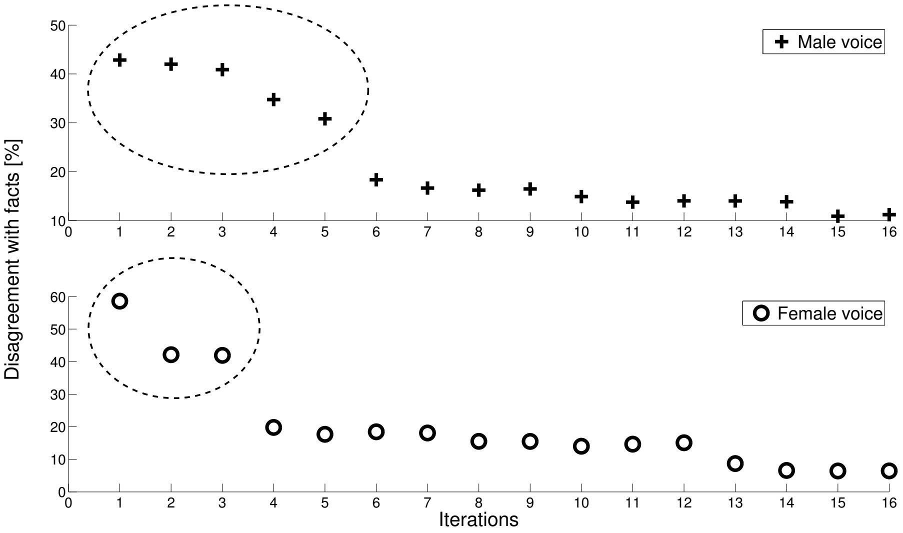


Fig. 2.3: Metric based on the listeners' *facts* disagreement rate.

The crosses and the circles grouped by the dashed lines represent a DISAGR value (2.2) of the worst agreeing listeners in each iteration.

2. Collect the *facts* based on the ratings given by all remaining listeners.
3. Calculate the disagreement score (2.2) of the excluded listener.
4. Return ratings of the excluded listener to the set of the ratings. Exclude ratings of another listener.
5. Repeat steps 2, 3 and 4 until all listeners are analyzed.
6. Find a listener with the highest disagreement score, discard her/his ratings and continue with the next iteration.

Correlation with Mean Opinion Score

Another metric, which is partly related to the analysis of the distribution of the listeners' ratings depicted in Fig. 2.2, is based on analyzing correlations of individual listeners with the group mean opinion score (MOS) given on the five-point scale. It is again possible to apply an iterative procedure, and in each iteration identify the least correlating listener, remove her/him from the

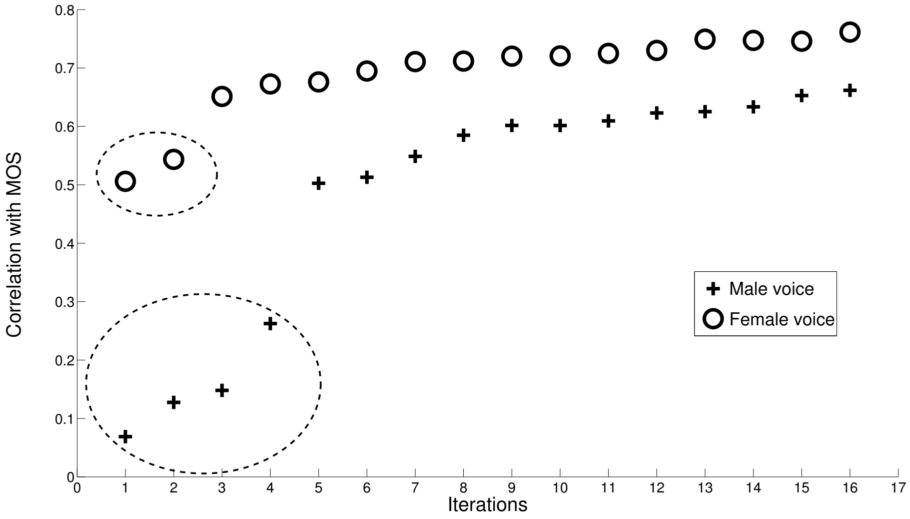


Fig. 2.4: Metric based on the listeners' correlations with the group MOS.

The crosses and the circles grouped by the dashed lines show listeners that were identified as deviating by the correlation analysis.

set of the listeners, and recalculate the MOS value. This process is illustrated in Fig. 2.4 showing that some deviating listeners can again be identified.

The steps of one iteration of the correlation analysis can be summarized as follows:

1. Exclude ratings of the first listener in the set.
2. Calculate MOS for each sentence given ratings by all remaining listeners.
3. Calculate the correlation with MOS of the excluded listener.
4. Return ratings of the excluded listener to the set of the ratings. Exclude ratings of another listener.
5. Repeat steps 2, 3 and 4 until all listeners are analyzed.
6. Find a listener with the lowest correlation with the MOS, discard her/his ratings and continue with the next iteration.

2.5 Preliminary Listening Test

2.5.1 Motivation

There were two main objectives for the preliminary listening test. First, as already mentioned in Sec. 2.2, we wanted to verify the feasibility of the proposed *half-sentence method* (see Sec. 2.2). Second, we wanted to make sure that data rich enough for the evaluation of the concatenation cost functions can be collected. To meet the objectives, using the female voice data was decided to be sufficient.

2.5.2 Test Stimuli

Types of Sentences

In addition to sentences synthesized by the *half-sentence method* (henceforth referred to as HS-ALL), another set of sentences was created. This set contained sentences in which the middle CVC words were synthesized using diphones taken from a diphone inventory created from the whole set of the recorded sentences. This set will be referred to as DI-ALL.

The difference between the HS-ALL and the DI-ALL sets was that the latter one contained sentences with three concatenation points in the middle word in contrast to the one concatenation point contained per sentence in the HS-ALL sentences. The inclusion of the DI-ALL set was motivated by the question whether restricting the prosodic content of the concatenation points by cutting words being placed in the same prosodic environment, i.e. the middle words of the three-word declarative sentences, was not limiting too much the range of possible discontinuities. The diphones were expected to represent richer prosodic features, e.g. a sentence final intonation or stressed/unstressed syllables, and thus to bring more audible discontinuities.

To lower the impact of the concatenation points in a close neighbourhood of the concatenation points of our interest, i.e. the concatenation points in the

Tab. 2.2: Subsets of sentences contained in the preliminary listening test.

Set	Description	Num.
HS-SEL	*half-sentence method* sentences	90
DI-SEL	mid-words synthesized using diphones	90
ran	random selection of sentences from HS-ALL and DI-ALL	15
nat	original recordings	5
rev	revision sentences	10
Total		210

middle of the vowels, spectrograms and oscilograms of all DI-ALL sentences were manually checked, and sentences containing some visible discontinuities at the surrounding concatenation points were removed.

Selection of Sentences

Since the number of sentences in both HS-ALL and DI-ALL sets was enormous, a selection of a subset of the sentences to be included into the preliminary listening test stimuli was needed. For that purpose, two approaches were used. First, the discontinuity metric based on the LSM approach described in Sec. 1.4.3 was implemented using three extraction window lengths ($K = 3, 4, 5$ pitch periods). Second, the discontinuities at the concatenation points were measured using the Euclidean distance between MFCC vectors extracted from the neighbouring frames of the concatenation points, three different window lengths were used for the feature extraction—10, 20 and 30 ms.

By each of these approaches, 15 best and 15 worst concatenations were found in both HS-ALL and DI-ALL sets. The resulting selection contained 180 sentences (6 x 2 x 15). These sentences will henceforth be denoted HS-SEL and DI-SEL, respectively. In addition, 15 randomly selected sentences were included (mixture of both DI and HS). For the reliability analysis reasons, 5 natural and 10 revision sentences were added. In Tab. 2.2 the counts of the selected sentences are summarized.

2.5.3 Procedure

The listening test was conducted following the procedure described in Sec. 2.3.2.

2.5.4 Evaluation and Conclusions

Reliability Analysis of Listeners

The total number of listeners who participated in the preliminary listening test was 19. In line with the procedures described in Sec. 2.4, the reliability analysis was performed. Based on the analysis results, two listeners were excluded due to obtaining low scores (-0.201 and -0.140) for rating natural and revision sentences, and additional two listeners due to the low correlation (0.24, 0.51) with the group MOS. The average score of the remaining listeners with respect to the ratings of the natural and revision sentences was -0.008, and the average correlation with the group MOS was 0.76, which was a considerably better result than reported in [Don01] where the average per-listener correlation with the group MOS was reported to be 0.49. This gives us a confidence that the test stimuli created by the *half-sentence method* can be consistently rated by non-expert listeners.

Collection of Facts

The next step of the evaluation was the collection of *facts* as defined in Sec. 2.4.4. In the HS-SEL set, including the HS random sentences, there have been found 39 *facts*—14 continuous and 25 discontinuous—, which represent approximately 42 % of the HS data. In the DI-SEL set, including DI random sentences, 9 continuous *facts* have been collected. Before any discontinuous *facts* could be found, it had to be ensured that none of the perceived discontinuities were due to poor quality of the joins surrounding the mid-vowel joins of our interest.

Tab. 2.3: Summary of *facts* collected in the preliminary listening test.

Set	**Continuous**	**Discontinuous**	**All**
`HS-SEL + ran (HS-ALL)`	14	25	39
`DI-SEL + ran (DI-ALL)`	9	17	26
Total	23	42	65

This was, of course, not a simple matter as it was not known how to estimate this impact. As already mentioned above, all DI-SEL sentences had been visually checked before their inclusion into the listening test stimuli. During the evaluation of the results, energy and pitch differences at all concatenation points were calculated, and used as an additional indicator.

Since human loudness perception does not scale linearly with the intensity of a signal, the energy differences at the concatenation points were measured on a logarithmic scale (dB). The same applies to the perception of pitch, for which the Mel scale was used. The maximum acceptable level of differences in both energy and pitch was found using all concatenation points in the sentences that were found to be the continuous *facts*. These values were then applied as thresholds for analyzing the neighbouring joins in the DI sentences. Upon removal of sentences containing neighbouring joins exceeding the found thresholds, 17 discontinuous *facts* have been identified in the DI-SEL set, including random DI sentences.

The summary of collected *facts* is presented in Tab. 2.3. It can be seen that the ratios of continuous and discontinuous *facts* are almost the same in both HS and DI sets. Interestingly enough, more *facts* were found in the HS set. The obtained results again show that the *half-sentenced method* is useful for preparing test stimuli that can be used for our purposes.

Energy and Pitch Differences at Concatenation Points

Since the objective of our work is to find a concatenation cost function that would reliably measure the discontinuities in spectral envelopes, pitch and

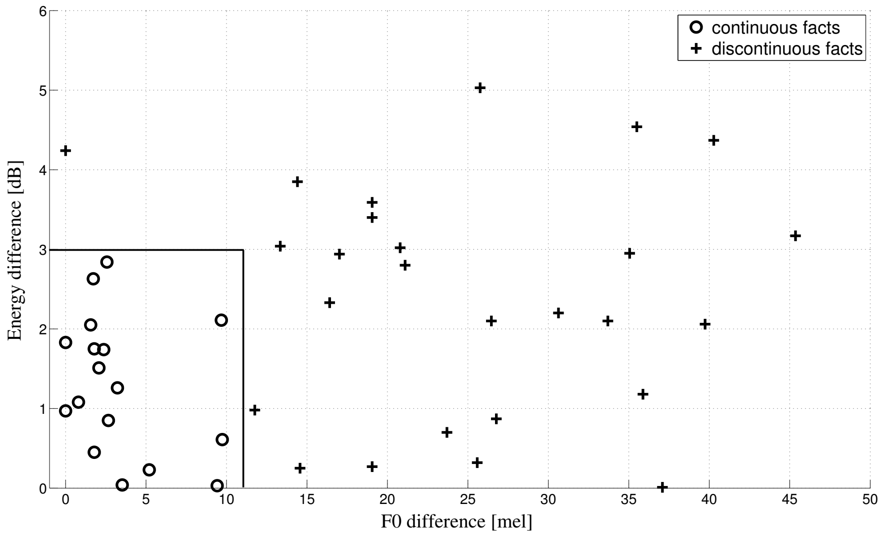


Fig. 2.5: Distribution of continuous and discontinuous *facts* in a $\Delta F0 \times \Delta Energy$ plain.

> *The continuous facts can be separated from the discontinuous ones in the* $\Delta F0 \times \Delta Energy$ *plain, which suggests that there are no sentences in which the perceived discontinuities could be attributed solely to the spectral envelope mismatches.*

energy at the concatenation points, assuming that they exist [Dut08], it is necessary to analyze whether the data collected in the listening tests contain sentences, in which discontinuities could also be attributed to mismatches in the spectrum.

In order to get some insight into possible impacts of the discontinuities in pitch and energy, a simple static difference analysis was conducted. The result of this analysis is depicted in Fig. 2.5, which shows that the discontinuous *facts* are linearly separable from the continuous ones in the $\Delta F0 \times \Delta Energy$ plain. The differences were measured using one pitch period long segments on both sides of joins as illustrated in Fig. 1.2.

There were no sentences found in the preliminary listening test, in which the discontinuity in spectral envelopes could be considered to be the only source of audible artifacts or, more precisely, that discontinuities would only be measurable in the signal spectrum. Obviously, there were cases where some considerable spectral mismatches at the concatenation points were observed. In all such cases was however the spectral mismatch accompanied by either a pitch or an energy mismatch.

Learned Lesson

As stated above, one of the key questions we wanted to address in the preliminary listening test was whether the concatenation points present in the sentences synthesized using the *half-sentence method* (see Sec. 2.2) contain enough discontinuities. Based on the obtained results summarized in Tab. 2.3, it could be concluded that this approach is usable as the number of perceived discontinuities follows the same trend in both HS and DI sentences. No statistically significant difference can obviously be found between the two approaches ($\chi^2 = 0.01$, $p = 0.916$).

It was also found out that the listeners were able to rate the sentences consistently as the average correlation between a per-listener ratings and the group MOS was 0.76. On the other hand, the ratio of *facts* with respect to the total number of sentences included in the listening test was only 42 %, which suggests that there were many items for which the listeners were in doubt if there is a discontinuity present or not. This number was one of the findings of the preliminary test which were used during the design of the large scale listening tests described in the next section.

Another very important finding was that all discontinuous *facts* were linearly separable from the continuous ones in the $\Delta F0 \times \Delta Energy$ plain. Therefore, more stress has to be put on the selection of sentences to be included into the listening test stimuli in order to find sentences in which the spectral envelope mismatches could be considered to be the major source of perceived discontinuity. This is crucial for the design of the spectral component of the concatenation cost functions.

2.6 Large Scale Listening Tests

2.6.1 Introduction

The objective of the large scale listening tests was to collect enough data that could be used for the design of a concatenation cost function and related experiments.

Since there were two main voices implemented within the framework of our TTS system [MRTT04] at the time when our work on the design of a new concatenation cost function was started, recordings of the voice talents of these two voices were used for our experiments. One of the speakers is female, the other one male. For each speaker, a listening test was prepared as described in the following subsections in order to collect perceptually annotated data.

2.6.2 Test Stimuli

A large number of synthesized sentences can be obtained from our recorded data using the *half-sentence method*. Based on the results obtained in the preliminary listening test, the method for selecting sentences to be included into the listening tests was revised. Alternative methods summarized in Tab. 2.4 were proposed. The goal was to be able to collect continuous and discontinuous *facts* that would be mixed when displayed in the $\Delta F0 \times \Delta Energy$ plain.

The subsets `f0B`, `enB` and `efB` were included to confirm that large differences in pitch and energy at concatenation points are a significant source of perceived discontinuities. For the selection of items of the `f0B` subset, the upper limit for the difference in energy was set to 1dB and sentences with maximum static $F0$ difference at concatenation points were selected. Similarly, the `enB` subset contained sentences where the difference in pitch was less than 10 mels while the difference in energy was maximum. The subset `efB` was composed of sentences of the largest distance from the origin in the $\Delta F0 \times \Delta Energy$ plain using the Euclidean distance. Since large measured

Tab. 2.4: Subsets of sentences contained in the large scale listening tests for each voice.

Set	Description	Num.
f0B	large pitch discontinuity and continuous energy transition	150
enB	large energy discontinuity and continuous pitch transition	150
efB	large pitch discontinuity and large energy discontinuity	150
efS	continuous energy and pitch transition	75
mfS	small pitch and energy difference + small MFCC distance	75
beS	small pitch and energy difference + small LSM distance	75
mfB	small pitch and energy difference + large MFCC distance	225
beB	small pitch and energy difference + large LSM distance	225
ran	random selection	135
nat	original recordings	15
rev	revision sentences	15
dbl	same source and target left and right consonantal contexts	20
Total		1310

differences in pitch and energy at the concatenation points often appear due to phonetic segmentation and/or pitch marking errors, all candidate sentences were checked manually, and the erroneous sentences were excluded.

In contrast to the preliminary listening test, more stress was put on sentences of smooth pitch and energy transitions at concatenation points, i.e. the subsets efS, mfS, beS, mfB and beB. The subset efS consisted of sentences of the smallest Euclidean distance from the origin in the $\Delta F0 \times \Delta Energy$ plain. For the selection of sentences included into the other four subsets, all sentences were ranked according to their Euclidean distance from the origin in the $\Delta F0 \times \Delta Energy$ plain, and only one third of the closest ones were taken into consideration.

The MFCC based distance was calculated as the Euclidean distance between two standard 12-dimensional MFCC vectors[4] characterizing the left and the right one pitch period long segments of the boundary region, respectively. The calculation of the LSM based distance was done as described in Sec. 1.4.3

[4] The first coefficient representing energy was not used.

using the SVD of the order 10 and the length of the extraction window set as $K = 3$ [Bel04].

In addition, a subset of natural sentences `nat` and a subset of randomly selected sentences `ran` were included. For the purposes of listeners reliability analysis, the subset `rev` containing revision sentences, i.e. sentences included twice into the listening tests stimuli, was added. The set of the listening test stimuli was completed by a subset of sentences that contained concatenation points but both the left and the right consonantal contexts were the same in the originally recorded and in the synthesized sentences. This subset of sentences is referred to as `dbl`.

The subsets `f0B`, `enB` and `efB` were expected to lead to discontinuous *facts*, the subsets `efS`, `mfS`, `beS`, `nat` and `dbl` to continuous *facts*, and the sets `mfB` and `beB` rather to the discontinuous *facts*.

The total number of sentences presented to listeners in each listening test was 1310. Note that the sentences themselves were not the same for both voices as the selection depended on the actual values measured for the synthesized candidates. The words containing the concatenation points (three words per vowel) and the number of sentences in each subset were however the same (see Tab. 2.4).

2.6.3 Procedure

Similarly to the preliminary listening test, the procedure described in Sec. 2.3.2 was used.

In addition, based on the lessons learned presented in [Ben05] and also on the feedback, which was collected in our preliminary listening test, the listeners were provided with a couple of examples of discontinuities to help them with calibration for the five-point scale. They were instructed to undergo the calibration phase before starting each listening test, and also after all breaks they made. It was allowed to listen to the calibration sentences at any time during the listening tests.

Tab. 2.5: Evaluation of listeners reliability.

	Female Voice	**Male Voice**
Number of listeners	27	29
Removed listeners	6	9
Average Correlation	0.75	0.67
Worst Correlation	0.67	0.51
Average Disagreement	14.58	12.98
Largest Disagreement	17.81	15.99

2.6.4 Evaluation

Evaluation of Listeners

There were 29 and 27 participants who finished the male and the female voice listening tests, respectively. Upon conducting the reliability analysis as described in Sec. 2.4, all listeners were ranked according to the obtained scores, and it was decided to remove 9 and 6 participants of the tests, respectively. The remaining participants were found to provide reasonably coherent ratings in terms of the correlations and also the *fact* disagreement rates. The summary of the listeners' reliability analysis is given in Tab. 2.5.

Collection of Facts

Using the ratings of all remaining listeners, the sets of *facts* were again collected using the definition (2.1). For the male voice, nicely balanced distributions of continuous and discontinuous *facts* were obtained for the vowels /a/,/e/ and /i/. The number of discontinuous *facts* was comparatively higher for the vowels /u/ (in line with the results presented in [KV01]) and /o/. For the female voice, more *facts* were collected for all vowels, which can be mainly attributed to the collection of discontinuous *facts*, especially for the vowels /a/, /e/ and /o/. As a matter of fact, very few continuous sentences were found for the female voice vowel /o/. The percentages and the counts of the

Tab. 2.6: Comparison of *facts* collections for the two voices.

	Female Voice	**Male Voice**
Number of *facts*	887	494
Ratio of *facts* [%]	67.71	37.71
Continuous *facts*	99	162
Ratio of continuous [%]	11.16	32.79
Discontinuous *facts*	788	332
Ratio of discontinuous [%]	88.84	67.21

Tab. 2.7: Comparison of *facts* collections across vowels—female voice.

	/a/	/e/	/i/	/o/	/u/
Number of *facts*	182	197	140	204	164
Ratio of *facts* [%]	20.52	22.21	15.78	23.00	18.49
Continuous *facts*	20	27	26	7	19
Ratio of continuous [%]	10.99	13.71	18.57	3.43	11.59
Discontinuous *facts*	162	170	114	197	145
Ratio of discontinuous [%]	89.01	86.29	81.43	96.57	88.41

Tab. 2.8: Comparison of *facts* collections across vowels—male voice.

	/a/	/e/	/i/	/o/	/u/
Number of *facts*	65	89	100	105	135
Ratio of *facts* [%]	13.16	18.02	20.24	21.26	27.33
Continuous *facts*	38	42	40	23	19
Ratio of continuous [%]	58.46	47.19	40.00	21.90	14.07
Discontinuous *facts*	27	47	60	82	116
Ratio of discontinuous [%]	41.54	52.81	60.00	78.10	85.93

collected *facts* are summarized in Tab.2.6-2.8. The *facts* are also graphically presented, sorted by vowels, in Fig 2.6.

Distribution of Facts in Test Stimuli Subsets

As described in Sec. 2.5.2, various subsets of sentences were contained in the listening tests stimuli to gain some control over the listening tests results in terms of the *fact* counts without having any a priori knowledge of the distribu-

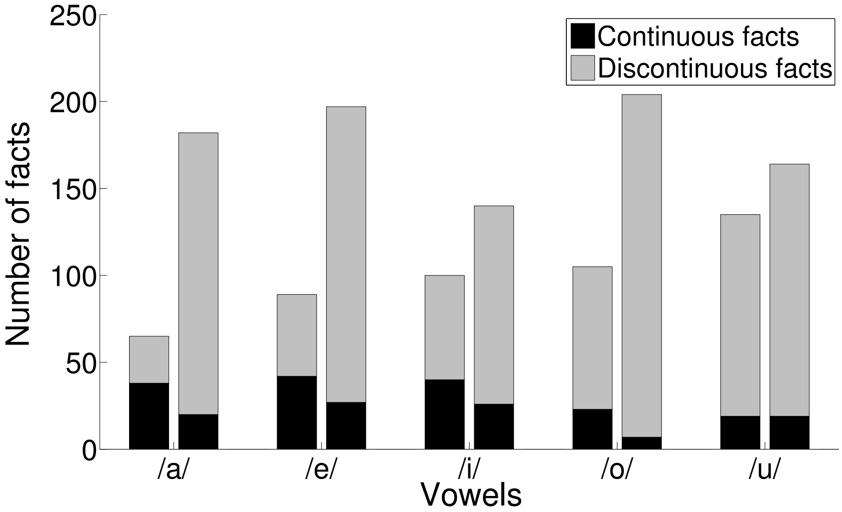


Fig. 2.6: Distribution of *facts* collected in the large scale listening tests.

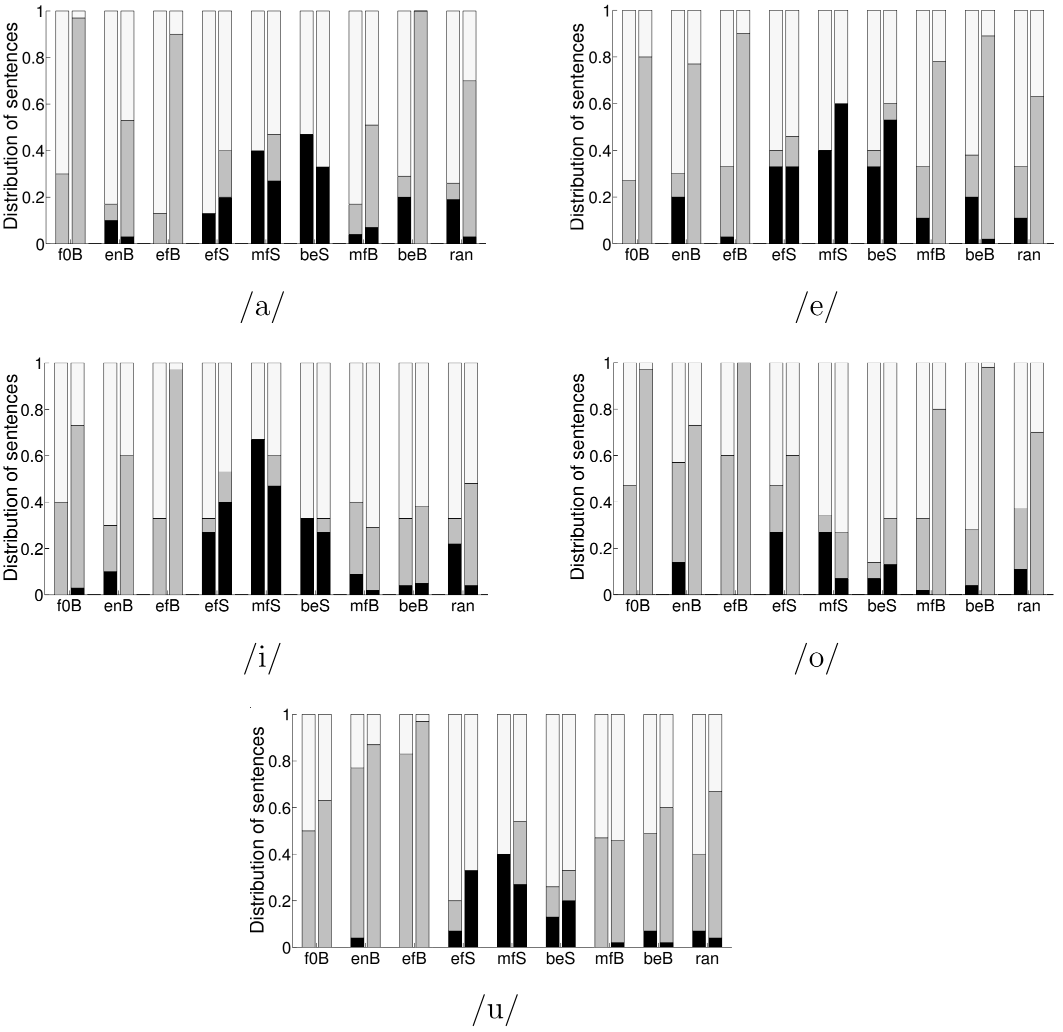
The facts collected in the listening tests are sorted by vowels—the left bar in each pair represents the male voice results.

tion of audible discontinuities in the synthesized data. The obtained results, depicted in Fig. 2.7, confirm that we only gained a limited control by introducing the selection methods that were used for the collection of the subsets of sentences. There are, however, some remarkable observations that can be made.

It is obvious upon closer inspection of the `ran` subsets of all vowels that the female voice synthesized data contained considerably larger amounts of audible discontinuities. Similar finding has already been presented in other studies comparing male and female voices, [Syr01] for instance.

The subsets `f0B`, `enB`, `efB`, `mfB` and `beB` show for the vowels /a/, /e/ and /o/ significantly smaller number of discontinuous *facts* for the male voice. In contrast, none of these subsets lead to different distributions when comparing the two voices for the vowel /u/, and the same can be said about the subsets `mfB` and `beB` for the vowel /i/.

If we turn next to the continuous *facts*, it is rather difficult to find any



/a/

/e/

/i/

/o/

/u/

Fig. 2.7: Distributions of *facts* collected in the large scale listening tests—test stimuli subsets.

> *Left bar in each pair represents results for the male voice. The continuous facts are represented by the black color, the gray color is used to show a proportion of the discontinuous facts. For instance, for the male voice vowel /a/, approximately 30 % of sentences were found to be discontinuous facts, and no sentence was rated as a continuous fact.*

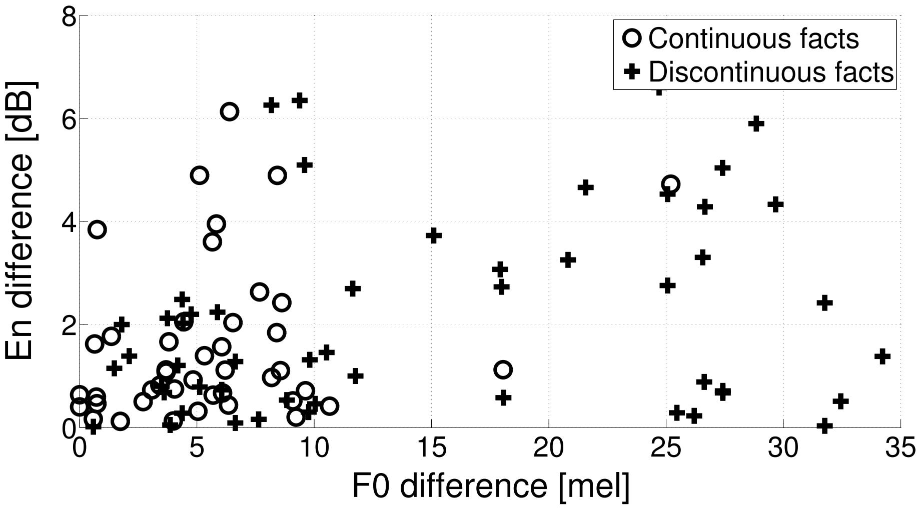


Fig. 2.8: Distribution of continuous and discontinuous *facts* in a $\Delta F0 \times \Delta Energy$ plain—vowel /e/, male voice.

In contrast to results obtained in the preliminary listening test, the continuous and the discontinuous facts are not linearly separable.

pattern across the vowels and the subsets. The differences in their quantities across the subsets overall compensate, which results in the comparable counts across the two voices as shown in Fig. 2.6. The only exception is the vowel /o/, for which surprisingly large number of the discontinuous *facts* can be found for the female voice in the subset `efS`.

Another observation that is worth mentioning when comparing the two voices is for the `beB` subset. Specifically, this subset generated very different results for the vowels /a/, /e/ and /o/, for which almost all sentences were rated as discontinuous *facts* for the female voice in contrast to the male voice.

Energy and Pitch Differences at Concatenation Points

Similarly to the preliminary listening test, the collected *facts* were analyzed from the perspective of their location in the $\Delta F0 \times \Delta Energy$ plain. As already mentioned in Sec. 2.5.4, one of the objectives of the large scale listening tests was to collect perceptual data, which would be exploitable also for measuring spectral envelope discontinuities. Fig. 2.8 shows that indeed, some sentences rated as discontinuous *facts* and having small static $F0$ and $Energy$ differences at concatenation points exist in the collected data. These sentences were expected to contain spectral discontinuities.

Another interesting point to see was how big the static discontinuities in pitch and energy at the concatenation points can be without being perceived by listeners. The locations of the continuous and the discontinuous *facts* for all voice/vowel combinations looked similarly to the one given in Fig. 2.8 showing the male voice vowel /e/. Interestingly enough, there have been some sentences that were assessed as continuous *facts* despite containing relatively large static differences in pitch and energy at the concatenation points.

It is noteworthy at this point that the plot only shows the static differences, i. e. differences calculated using one frame on each side of a concatenation point, and thus do not reflect the dynamics of the pitch and energy contours, which may also play an important role in the perception of discontinuities. This particular hypothesis will be addressed in Chapter 3 of this thesis.

Chapter 3

Role of F0 Discontinuities

3.1 Introduction

It is believed that discontinuities in $F0$ at concatenation points are the most important source of audible concatenation artifacts in voiced sounds. This has been confirmed formally [BSF05], and it is also supported by observations made during the collection of our experimental data as described in Sec. 2.6. It is however not known what a perceptual threshold for "affordable" static $F0$ differences at the concatenation points is. The role of pitch contour slopes has not been to our knowledge investigated either. Moreover, not all large $F0$ differences seem to necessarily lead to audible discontinuities as shown in Fig. 2.8.

In order to analyze the impacts of pitch contours (including their slopes and static differences at the concatenation points) on the quality of concatenations in vowels, a set of experiments was conducted. These experiments and their results are described in the following sections. The goal was also to clearly identify sentences in which audible discontinuities could be attributed to other aspects than $F0$.

3.2 Pitch Contours as Predictors of Concatenation Artifacts

3.2.1 Motivation and Approach

The experiment presented in this section was aimed at answering the question of how much information is contained in pitch contours (their slopes, shapes, static differences, etc.) with respect to discontinuities perceived by listeners.

Since analyzing $F0$ contours in concatenation areas and defining expert rules manually would be impractical, the task was formulated as a binary statistical classification problem using pitch contours extracted from the vicinity of concatenation points and/or their parametrization as predictors. Four different sets of $F0$ based predictors, described in Sec. 3.2.2, were defined. Support Vector Machines (SVMs) were chosen for training a classification model due to their proven good performance for different classification tasks.

The hypothesis under question was that concatenating incoherent[1] $F0$ contours lead to perceived discontinuities, which should be learned by the classifier, whereas coherent $F0$ contours are not sufficient condition for perceptually smooth concatenations (due to expected existence of other than $F0$ related concatenation artifacts), which should be decreasing the classifiers' sensitivity.

The data used in the experiment were composed of *facts* collected in the large scale listening tests described in Sec. 2.6. Since the continuous *facts* collected in the listening tests were for some vowels rather underrepresented compared to the discontinuous ones (see Fig. 2.6), it was decided to enrich the experimental data with some natural sentences in order to make them better balanced.

[1] Note that coherence does not necessarily mean at this point an exact match of concatenated contours but could also include cases when the static difference is large while slopes are similar.

3.2.2 Discontinuity Detection Predictors

Sets of Predictors

As a preparation for defining different sets of predictors, the original recorded sentences were pitch marked using the robust multi-phase pitch marking algorithm [LMT11]. Since no pitch smoothing method was applied during synthesis by the *half-sentence method*, pitch marks remained preserved in the synthesized sentences. Using the pitch mark sequences, fine-grained $F0$ contours could be calculated, and the following sets of predictors were created:

- `Reg` : $[M_{-1}, M_1, K_L, K_R]$
- `SReg` : $[\hat{K}_L, \hat{Q}_L, \hat{K}_R, \hat{Q}_R]$,
- `Syn` : $[L_{-4} \dots L_{-1}, P_1 \dots P_4]$
- `Nat` : $[L_{-4} \dots L_{-1}, L_1 \dots L_4, P_{-4} \dots P_{-1}, P_1 \dots P_4]$,

where L_i and P_i represent $i-$th point of natural $F0$ contours pitch synchronously extracted from the vicinity of concatenation points from vowels that were concatenated. The notation is depicted in Fig. 3.1. Values M_{-1} and M_1 were calculated as:

$$M_{-1} = (L_{-2} + L_{-1})/2$$

$$M_1 = (P_1 + P_2)/2$$

Values K_L and K_R are the slopes of linear regression lines fitted to the left and right natural $F0$ contours, respectively, and the pairs $\hat{K}_L$, $\hat{Q}_L$ ($\hat{K}_R$, $\hat{Q}_R$) were obtained as parameters of linear regression lines fitted to sequences $[L_{-2} \dots L_2]$ ($[P_{-2} \dots P_2]$), which were first smoothed by a median filter.

Rationale

The `Reg` set was included to address the assumption that static differences in pitch at concatenation points together with slopes of concatenated $F0$ contours represent the key predictors of audible $F0$ discontinuities.

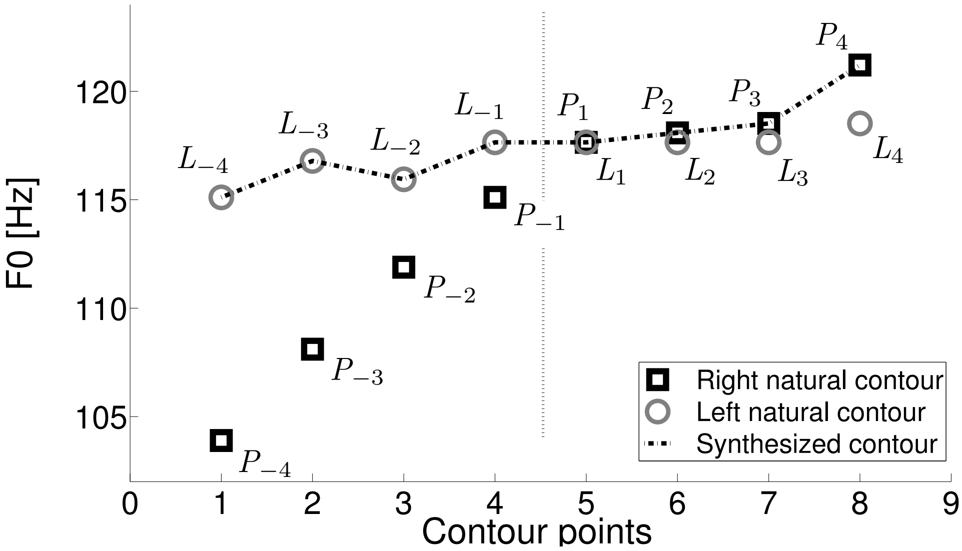


Fig. 3.1: Annotation scheme of $F0$ contours used for defining discontinuity detection predictors

> *As an example, let* $[L_{-4} \dots L_4]$ *be an* $F0$ *contour extracted from a central part of a vowel /a/ in a word /t_Sak/ and* $[P_{-4} \dots P_4]$ *a contour of /a/ in a word /mas/. Then, the sequence* $[L_{-4} \dots L_{-1}, P_1 \dots P_4]$ *represents a central part of a concatenated* $F0$ *contour of a word created as /t_Sa-as/.*

Since the estimated slopes of the $F0$ contours may be significantly affected by gross pitch marking errors, the SReg set was included. Considering the results of evaluations of the accuracy of the pitch marking algorithm [LMT11], no big differences were expected when comparing the performance of the classifiers trained on the Reg and the SReg sets of predictors.

The Syn set was purely composed of synthesized $F0$ contours. These contours do not contain any information about points of the $F0$ contours following the left part and preceding the right part of a synthesized vowel in the natural data. Since no pitch smoothing was applied during concatenating halves of the recorded sentences, there might have been considerable $F0$ jumps at the concatenation points. At the same time, the synthetic $F0$ contours may also appear to be very smooth, even in cases where the original natural contours have rather different slopes as shows the example depicted in Fig. 3.1.

To include a full description of the $F0$ contours of concatenated units within concatenation areas, the `Nat` set composed of both natural concatenated $F0$ contours extracted from recorded sentences from the vicinity of prospective concatenation points was added.

3.2.3 Training Classification Models

As suggested in [BHW10], first, a model using the linear kernel was trained. This can serve as a baseline, and can then be compared to the results obtained for a non-linear kernel model—the Gaussian (RBF) in our case.

To find the best SVM hyperparameters, a grid search using grid points distributed on a logarithmic scale was conducted. In the first step, we used a coarse grid to find promising regions, and then we further searched for better hyperparameters' values using a finer grid. The K-fold cross-validation technique[2] ($K = 5$), was used to estimate the classifiers' performance at each point on the grid.

3.2.4 Results

Linear Kernel Models

We turn first to the results of the classification using the linear kernel SVMs. The classifiers' performance rates in terms of accuracy (ACC), sensitivity (recall rate, $SENS$) and specificity ($SPEC$) averaged across all vowels are presented in Tab. 3.1. The sensitivity and the specificity were in our case defined as follows:

$$SENS = \frac{TRUE\_CONT}{TRUE\_CONT + FALSE\_DISCONT} \tag{3.1}$$

[2] Note that the cross-validation should help to prevent the overfitting problem, which is often a concern when training classification or regression models.

Tab. 3.1: Classification results—Linear kernel SVMs (average across all vowels)

	Male			Female		
Predictors	ACC	SENS	SPEC	ACC	SENS	SPEC
`Syn`	0.74	0.73	0.71	0.62	0.35	0.84
`Nat`	**0.79**	**0.87**	**0.68**	**0.72**	**0.76**	**0.69**
`Reg`	0.73	0.78	0.65	0.65	0.56	0.72
`SReg`	0.76	0.83	0.68	0.66	0.47	0.82

$$SPEC = \frac{TRUE\_DISCONT}{TRUE\_DISCONT + FALSE\_CONT}, \tag{3.2}$$

where *TRUE_CONT* is a number of continuous *facts* classified as such, *FALSE_DISCONT* is a number of discontinuous *facts* classified as continuous, *TRUE_DISCONT* is a number of correctly classified discontinuous *facts*, and *FALSE_CONT* is a number of continuous *facts* classified as discontinuous.

While sensitivity measures the proportion of correctly classified continuous *facts*, specificity gives the proportion of correctly identified discontinuous *facts*. These measures were calculated in order to get more insight into the performance of the classifiers as well as to address the hypothesis formulated in Sec. 3.2.1.

It can be seen that the accuracy of the SVMs using the linear kernel is not very high. It is, however, a promising result, taking into account the difficulty of the classification task. The classifiers performed worse on the female voice data than on the male voice data. Regarding the different sets of predictors, the `Nat` set gives the best results. This observation may be attributed to the fact that using the whole $F0$ contours increases the variance in the data, which may help the linear kernel SVMs to find better separation between the two classes. Further discussion on the obtained results will be presented in Sec. 3.2.6.

Tab. 3.2: Classification results—Gaussian kernel SVMs (average across all vowels)

	Male			Female		
Predictors	ACC	SENS	SPEC	ACC	SENS	SPEC
`Syn`	0.89	0.92	0.86	0.90	0.87	0.91
`Nat`	**0.93**	**0.96**	**0.91**	**0.92**	**0.95**	**0.91**
`Reg`	0.91	0.93	0.88	0.92	0.91	0.93
`SReg`	0.90	0.90	0.90	0.92	0.93	0.90

Gaussian Kernel Models

Having obtained the results for the linear kernel SVMs, the question was how much the models can be improved by introducing the non-linear kernel. The values presented in Tab. 3.2 show that all sets of predictors lead to comparatively higher performance rates.

By contrast, almost no difference was found between the averaged results for the male and the female voice data. The `Nat` predictors lead to the best classification results, and the `Syn` set shows, similarly to the linear kernel SVMs, the worst results. This suggests that the knowledge of the whole concatenated $F0$ contours is beneficial.

If we look at the variance of the classifiers' performance across different vowels presented in Fig. 3.2, it can be seen that the accuracy of the classification was lower for the vowel /i/, especially for the female speaker.

Another interesting comparison that can be made, is to put the obtained results into relation to the disagreement scores of listeners (eq. 2.2) as these in fact represent the error rates of the listeners performing the same classification task as the SVM models. The disagreement score can simply be converted to the agreement score, i.e. the classification accuracy:

$$AGR\_SCORE_i = (100 - DISAGR_i)/100 \tag{3.3}$$

The agreement scores (3.3) of the three least agreeing listeners for each voice,

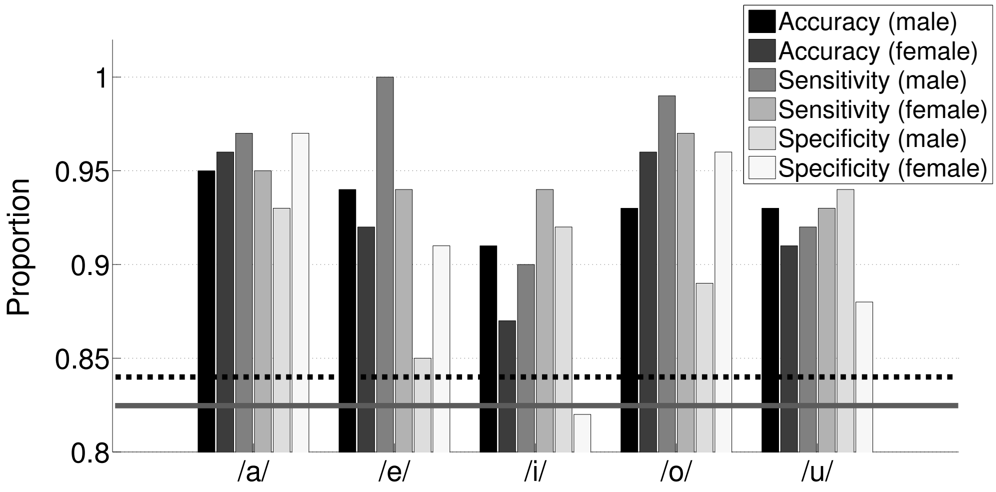


Fig. 3.2: Comparison of classification results across all vowels and both speakers—`Nat` set of predictors.

The dotted black and solid gray lines show the lowest agreement scores (3.3) obtained from the listeners participating in the large scale listening tests for the male and the female voice, respectively.

Tab. 3.3: Agreements scores (3.3) of the three least agreeing listeners participating in the large scale listening tests.

	Male	**Female**
`List1`	0.84	0.82
`List2`	0.87	0.83
`List3`	0.88	0.84

which may serve as a reference for the evaluation of the classifiers' performance, are summarized in Tab. 3.3. The score of the least agreeing listener in each test is also depicted in Fig. 3.2.

Summary of Models' Hyperparameters

For completeness' sake, we present in Tab. 3.4 the values of the SVM models' hyperparameters obtained during training on the `Nat` set of predictors. It can be seen that the hyperparameter values of most of the models are relatively

Tab. 3.4: Hyperparameter values of the SVM models.

Speaker/Vowel		**Linear**	RBF	
		C	**C**	γ
Male	/a/	0.0385	5.6569	0.0412
	/e/	1.0718	1.7411	0.2679
	/i/	1.6245	0.5743	0.1340
	/o/	8.0000	10.556	0.1649
	/u/	2.6390	2.8284	0.3536
Female	/a/	8.0000	6.0629	1.3195
	/e/	8.5742	6.9644	0.7579
	/i/	2.4623	64.000	0.0292
	/o/	24.252	3.2490	1.7411
	/u/	1.4142	6.4980	0.4061

small suggesting that the models should be capable of generalization [BHW10]. The exception is the Gaussian kernel SVM model for the female voice vowel /i/, which tends to overfit. That partly explains the lower accuracy estimate obtained by the cross-validation.

3.2.5 Verification of the Models

Objective Verification

Since the results presented in the previous section are very promising, especially taking into account the extreme difficulty of predicting audible concatenation artifacts, there was a suspicion that the models are overfitted to the training data. Generally speaking, overfitting should theoretically be avoided by performing the cross-validation during the training, and as mentioned above, our results were indeed obtained via the cross-validation procedure. Nevertheless, another possible source of over-fitting present in our data could be the repeating halves of sentences involved in the concatenations.

In order to investigate whether or not this source of overfitting plays a role, we conducted the following experiment. For each half of a sentence, we randomly generated a sequence of values to replace its original pitch synchronously

Tab. 3.5: Comparison of accuracies of the SVM models—real vs. random data

	Female		**Male**	
	Real	**Random**	**Real**	**Random**
/a/	0.95	0.83	0.96	0.54
/e/	0.94	0.58	0.92	0.54
/i/	0.91	0.63	0.87	0.54
/o/	0.93	0.54	0.96	0.54
/u/	0.93	0.54	0.91	0.54

extracted $F0$ contour. These random $F0$ contours were used instead of the sequences representing the real $F0$ contours in the training data. Using this approach, the repetitions of the $F0$ contours in the training data were kept but the contours were random. If the repetitions were the source of overfitting, there would not be a considerable drop in the performance of the models. The `Nat` set of predictors was used for this experiment. The comparison of accuracies of the models trained on the original and on the randomized data are presented in Tab. 3.5.

It can be seen from the obtained results that the models cannot be well trained on the training data consisting of the random $F0$ contours, which proves that they are not capable of learning the pure repetitions. The only subset of our data where the repetitions may play a role, albeit limited, is the vowel /a/ for the male voice. Note that for this particular vowel, we indeed had less training data, which probably makes the repetitions play a more important role. Nevertheless, even for this case, there is still a significant drop in the model's accuracy when comparing real and random predictors.

Subjective Verification

No additional listening tests dedicated to the subjective verification of the performance of the trained classification models were organized. Nevertheless, in Chapter 4 of this thesis where the effect of consonantal contexts on the quality of concatenations in vowels is investigated, the trained models are used to factor out the impact of $F0$. More details can be found in Sec.4.2.3.

The results of that analysis presented in Tab. 4.6-4.7 clearly show that most of the sentences classified by the models as containing continuous concatenations indeed do not contain any audible artifacts.

3.2.6 Discussion

Based on the assumption that concatenating coherent $F0$ contours is necessary but not sufficient condition of perceptually smooth concatenations (not applying any smoothing), and that concatenating incoherent $F0$ contours leads in most cases to perceptually discontinuous joins, the sensitivity of the classification models was expected to be comparatively lower than their specificity.

As can be seen from Tab. 3.2, our expectation was rather not supported by the actual measurements showing that different sets of predictors lead to different results. The `Nat` set, for which the highest classification accuracy was achieved, shows the opposite of what was originally expected. If we look more closely at Fig. 3.2, we can see that the sensitivity and the specificity rates may vary from vowel to vowel, and even inconsistently when comparing the two speakers.

This observation does not necessarily disconfirm the assumption that coherent $F0$ contours are the necessary condition for the perceptually smooth concatenations. Since the models were trained with respect to their accuracy, of which we believe quite robust cross-validation estimates were obtained, and the specificity and the sensitivity rates may to some extend vary depending on the randomization of the order of the training data, it suggests that there are some clusters of $F0$ contours, which are not well separable. It is then the matter of training, into which class these clusters are put. This results in the variance of the sensitivity and the specificity measures. These difficult clusters must however be rather non-dominant in our data since the models' accuracy remain high.

As a matter of fact, calculating confidence intervals for cross-validation estimates is considered to be a difficult problem. Nevertheless, if we look at the models' classification accuracy, assuming that the bias of its estimates is rather towards a poorer fit (which is believed to be true for cross-validation estimates), and make the comparison with the agreement scores listed in Tab. 3.3, which are slightly biased in the direction of higher values (due to the participation of each listener in the creation of the *facts*), we can see that the SVM classifiers perform very well, and the obtained high accuracy is clearly exceeding our expectations.

It is, however, important to mention at this point that the presented results are not meant to question the role of the spectral envelope and/or phase mismatches[3] in the perception of the concatenation discontinuities. They should rather suggest that the discontinuities are detectable with a high accuracy using the $F0$ contours as predictors, and this knowledge is beneficial for improving the concatenative speech synthesis.

3.2.7 Conclusions

We have presented the results of audible discontinuity detection task performed by the SVM classifiers trained on $F0$ contours extracted from the vicinity of concatenation points and/or their parametrization. The results show that the information contained in the contours is sufficient to detect a large number of audible concatenation discontinuities with a high accuracy falling into the range around 90 %, which is unquestionably a very good result. The presented results suggest that putting more stress on the $F0$ contours during unit selection and improving their modeling may be a promising way to improve the output of our TTS system, at least for vowels.

[3] Phase mismatches are actually shown in this work to play an important role in certain contexts. More details can be found in Sec. 4.2.6.

The Gaussian kernel SVMs were found to be giving better classification results than the linear kernel SVMs. The best classification accuracy was achieved using all points of the $F0$ contours extracted pitch synchronously from the vicinity of prospective concatenation points from both concatenated diphones. Nevertheless, the parametrization of the contours by linear regression (no matter if the contours are pre-smoothed or not) does not significantly decrease the models' accuracy. This can be beneficial for reducing the memory requirements of the TTS system. Using only the synthesized $F0$ contours is slightly inferior. This means that the information contained in the natural $F0$ contours spanning the concatenation areas in both concatenated diphones should be leveraged.

The results have also shown that the specificity and the sensitivity rates may vary across vowels from speaker to speaker, and also for different sets of predictors, which does not support, neither disconfirm, the assumption that concatenating coherent $F0$ contours is necessary but not sufficient condition of perceptually smooth concatenations.

3.3 Clustering of Pitch Contours

3.3.1 Motivation

Upon obtaining very promising results in the experiments with the SVM models, two questions arose: (1) What is the latent construct contained in the $F0$ contours which allows the models to distinguish continuous and discontinuous *facts*. (2) How can the information carried by the $F0$ contours be leveraged to improve the quality of the output of our TTS system.

Since the kernel trick, which is used by the SVM algorithm, allows for a linear separation of the observations in an inner product feature space of a high dimension without an explicit calculation of the mapping function from the original feature space, it is not a simple matter to understand the classifier's decisions and turn them into an expert knowledge. As a consequence, the latent constructs, we mention above, cannot unfortunately be simply seen. Therefore, the $F0$ contours of concatenated units need to be further analyzed by other means.

Also, integrating the SVMs directly into the TTS system is not completely straightforward due to a couple of reasons, including their computational heaviness or the current system's architecture. A promising approach could be implementing a caching mechanism for the concatenation costs similarly to [BMR99].

To get more insight into the importance of the similarity of concatenated $F0$ contours, a simple clustering experiment described in the following subsections was proposed. If the clustering proved to be a feasible way of identifying units that can be well concatenated, it could be more easily incorporated into our unit selection based TTS system than the SVM classifiers themselves. The results of clustering are also easier to interpret, which would help us to understand the $F0$ related cues that drive the listeners' discontinuity decisions.

3.3.2 Experiment Setup

Set of Observations

To maintain equal conditions, the set of observations to be clustered was given by the best predictors used in the classification experiment described in Sec. 3.2, i.e. the `Nat` set was used. As already described, this set contains points of the $F0$ contours pitch synchronously extracted from the recorded sentences from the areas around prospective concatenation points for each pair of concatenated diphones. The set of observations on a given scale was then formally defined using the notation shown in Fig. 3.1 as:

$$OBS = \{\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_M\} \cup \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N\}\,, \tag{3.4}$$

where $\mathbf{l}_i = [L_{-4} \ldots L_4]_i$, $\mathbf{p}_j = [P_{-4} \ldots P_4]_j$, and M, N are the counts of sentences that appeared in synthesized data as left and right halves, respectively.

To respect the non-linearity of human perception, we mapped the set of observations to two perceptually motivated scales—mel and semitone—in addition to the basic representation on the hertz scale.

Procedure

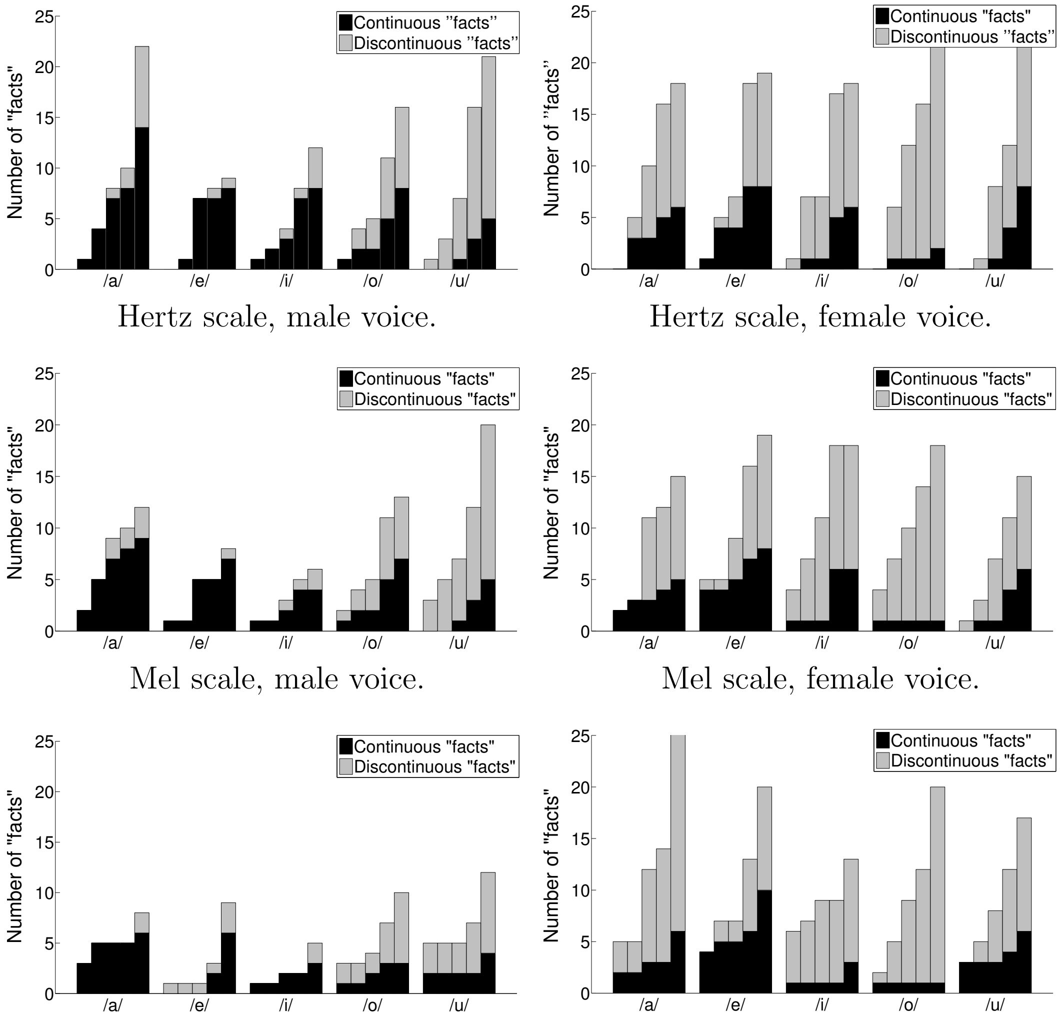
Upon building a clustering tree[4], different heights (diameters) were applied to cut it into different numbers of clusters. Within each cluster, the continuous and discontinuous *facts* were then counted. Following the hypothesis that proximity of the concatenated $F0$ contours is the key factor for discontinuity detection, the concatenations within small-diameter clusters should rather be continuous, and the larger the diameter gets the larger number of discontinuous concatenations should be appearing. In addition, most of the existing continuous *facts* were expected to be found within clusters, not across clusters.

3.3.3 Results - Within Cluster Concatenations

The results of the procedure described in the previous paragraph are presented in Fig. 3.3 showing distributions of *facts* within clusters of different diameters for all vowels and comparing different scales. It can be seen from the plots that our hypothesis was not confirmed by the obtained results, especially for the female voice and the male voice vowels /o/ and /u/.

Also for the other male voice vowels, the results were not very convincing taking into account the total number of continuous *facts* present in our data, and comparing this number with the counts of continuous *facts* present in the clusters. It is obvious from this comparison that there have been many continuous *facts*, which did not fall within the same clusters. This means

[4] The observations were clustered using DIvisive ANAlysis (DIANA) clustering technique and the Euclidean distance. This technique is described in Chapter 6 of [KR05].



Hertz scale, male voice.

Hertz scale, female voice.

Mel scale, male voice.

Mel scale, female voice.

Semitone scale, male voice.

Semitone scale, female voice.

Fig. 3.3: Distributions of *facts* sorted by voices and scales.

Each particular bar contains a distribution of facts for a given cutting height of the clustering tree. For each vowel, five cluster diameter values were used to plot the distributions. They are ranked in ascending order from left to right. The larger the diameter of the final clusters is the more concatenations within the clusters appear—including both continuous and discontinuous facts.

that the concatenated $F0$ contours are rather incoherent according to the results of the clustering procedure, but still the listeners did not perceive any discontinuity, and vice versa.

The same unfortunately holds true even for the perceptual scales—mel and semitone, which were expected to lead to better results.

3.3.4 Analysis of Concatenations across Clusters

Having obtained the results of the within cluster concatenations, the natural question was what concatenation quality can be seen when concatenating across different clusters. Obviously, there can be clusters, the combinations of which lead to rather continuous or rather discontinuous concatenations. The hypothesis was that if such clusters exist, it is not only the distance between their centers, but also the actual positions and shapes of the $F0$ contours of these clusters which would play an important role. The objective of this analysis was again the same, i.e. trying to understand what the latent constructs learned by the SVM models could be. The analysis was done for the cluster diameters in the range from 10 to 25 Hz, which corresponds to the diameters used in the hertz-scale plots shown in Fig. 3.3, and consisted in counting discontinuous concatenations which the sentences of individual clusters were involved in. The rationale for using different values of the diameters is discussed below in Sec. 3.3.5.

The results of the analysis are shown in Fig. 3.4-3.13. The plots show frequencies of concatenations the sentences from particular clusters were involved in, and how many of these concatenations were found to be the discontinuous *facts*. The numbered points in the plots represent individual clusters. The sentences of clusters lying on the diagonal lines were involved in discontinuous concatenations in all cases. The clusters between the diagonal line and the upper dashed line are those for which 75 % of concatenations were found discontinuous. The sentences from the clusters below the lower dashed lines lead to the discontinuous concatenations in less than 25 % of cases.

Despite the fact that many clusters are involved in very few concatenations, which leads to the high density of their positioning in the plots, some outlying clusters can clearly be identified for each vowel-speaker combination with the exception of the female vowel /o/ where the vast majority of concatenations were found to be discontinuous *facts* as already mentioned in Sec. 2.6.4.

The outlying clusters are of two types—continuous and discontinuous—and will be hereafter referred to as *critical clusters*. The *continuous* and *discontinuous critical clusters* are those which rather lead to continuous and discontinuous concatenations, respectively. They can be found in Fig. 3.4-3.13 above the upper and below the lower dashed lines. These two lines will henceforth be referred to as *criticality borders*.

3.3.5 Further Investigation of Critical Clusters

Motivation

Generally speaking, there are at least two possible reasons for the existence of the *critical clusters*. First, the consonantal contextual phonetic effects, which may play a role due to repeating occurrence of the same halves of sentences in the concatenations. Second, an outlying position of the $F0$ contours contained in these clusters, which may result in a difficulty to find perceptually appropriate counterparts for concatenating, or, contrariwise, a central position of the contours, which can be suitable for concatenating with any of the surrounding clusters. We keep the former as a subject to a detailed analysis, which will be described in Chapter 4, and only the latter point will be dealt with in this section.

Let us now get back to the reason for choosing different diameter values in the analysis of the across cluster concatenations described in the previous section. Obviously, the more restrictions are posed on the compactness of the clusters, i.e. the lower the cluster diameter value is set, the less contours are contained in each cluster, and the more likely a few phonetically important

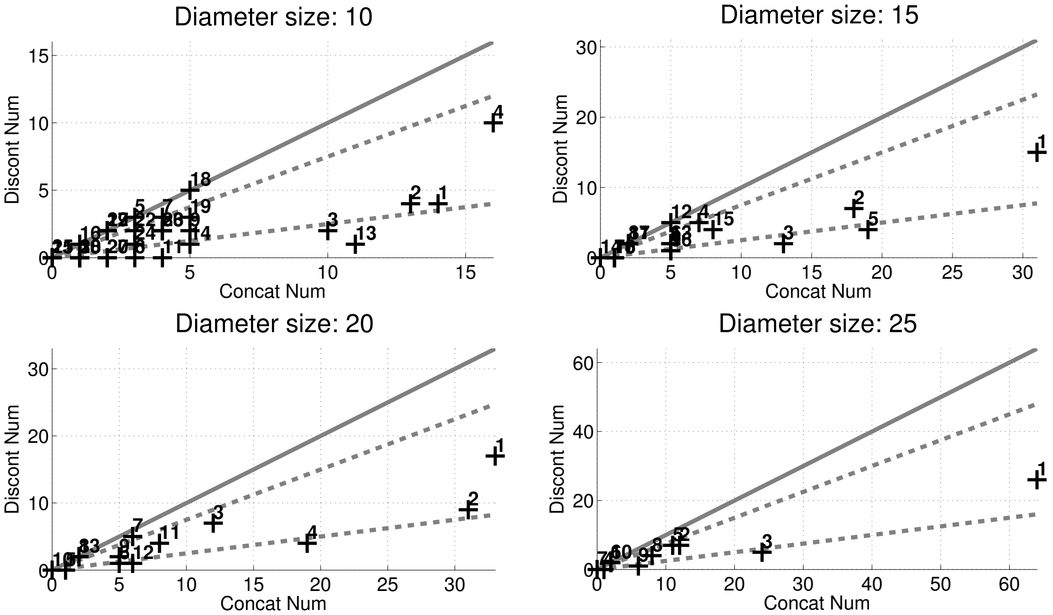


Fig. 3.4: Discontinuous concatenations per cluster - vowel /a/, male speaker

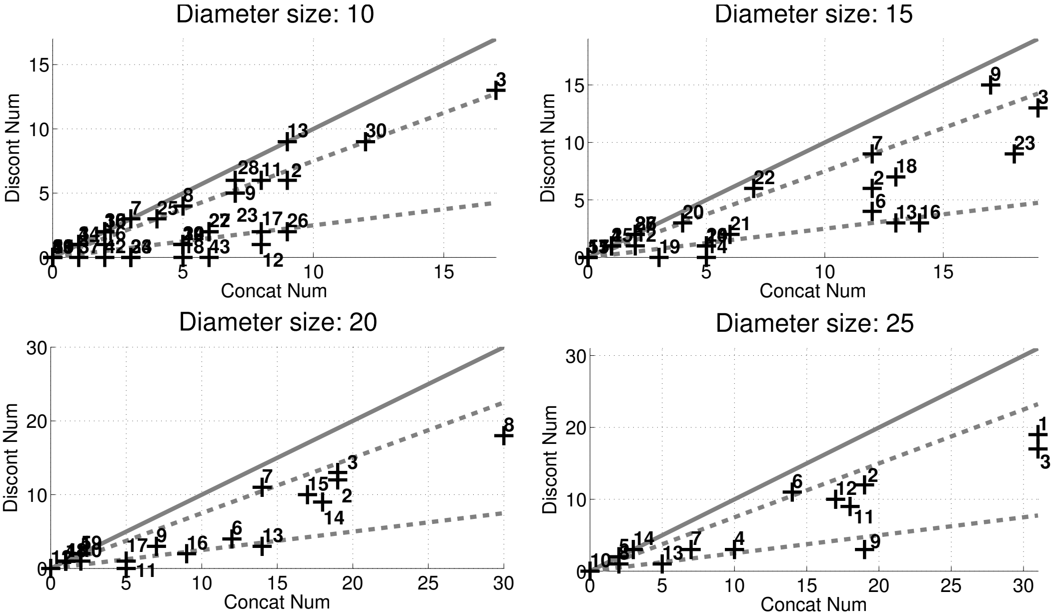


Fig. 3.5: Discontinuous concatenations per cluster - vowel /e/, male speaker

Clusters are represented as numbered points. The diagonal line represents 100 % of discontinuous concatenations. The dashed lines represent 75 % and 25 % of discontinuous concatenations, respectively. Cluster diameter values are in Hz.

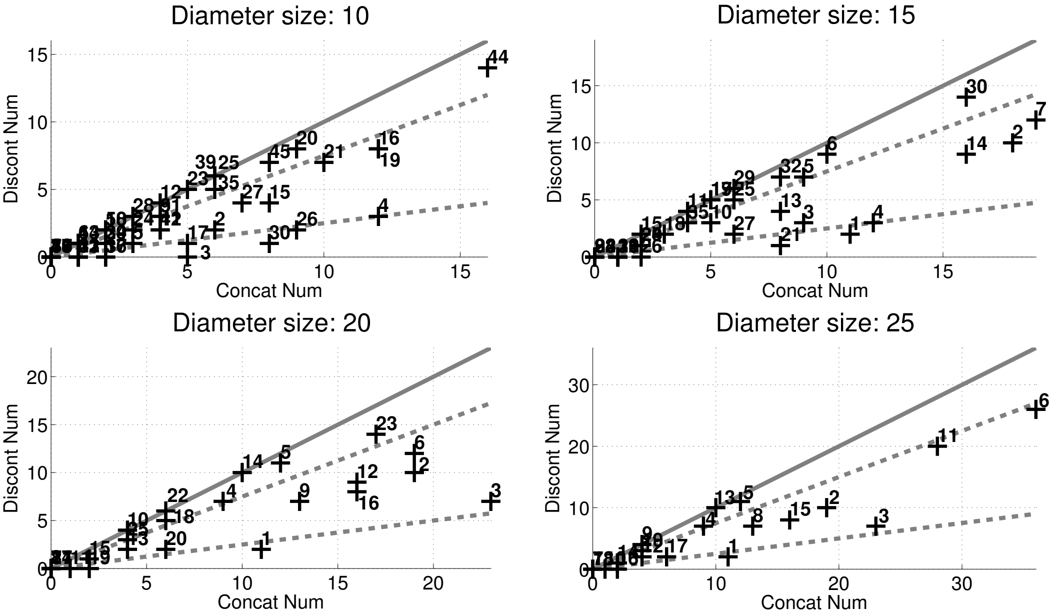


Fig. 3.6: Discontinuous concatenations per cluster - vowel /i/, male speaker

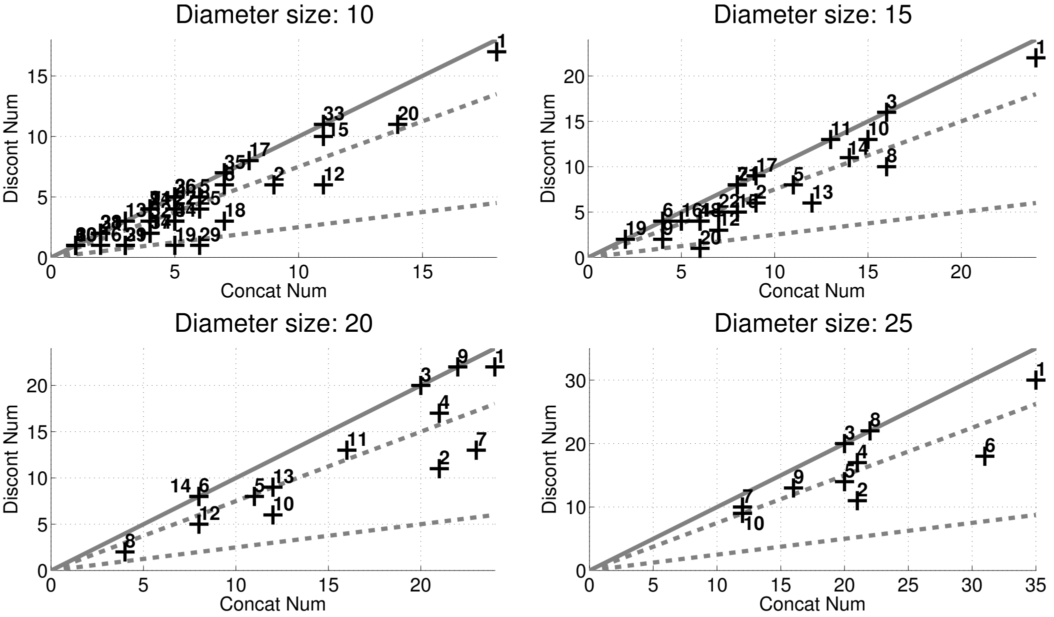


Fig. 3.7: Discontinuous concatenations per cluster - vowel /o/, male speaker

Clusters are represented as numbered points. The diagonal line represents 100 % of discontinuous concatenations. The dashed lines represent 75 % and 25 % of discontinuous concatenations, respectively. Cluster diameter values are in Hz.

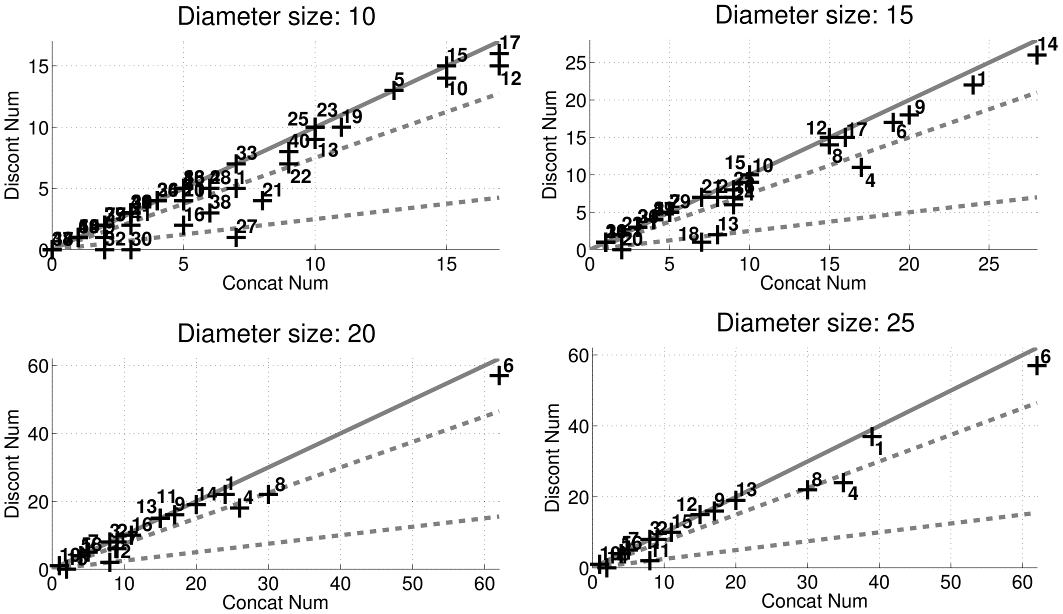


Fig. 3.8: Discontinuous concatenations per cluster - vowel /u/, male speaker

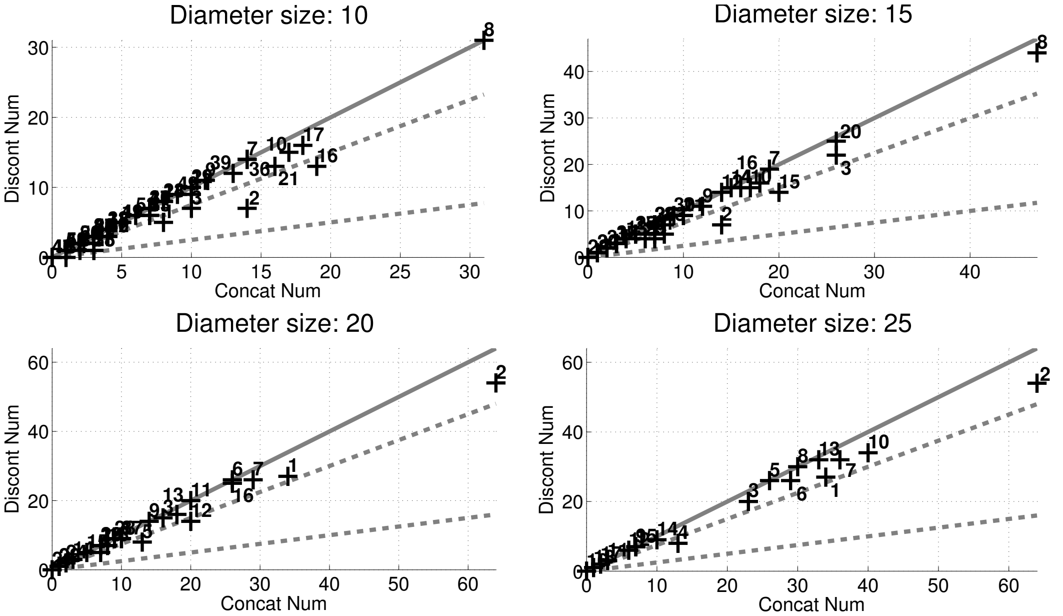


Fig. 3.9: Discontinuous concatenations per cluster - vowel /a/, female speaker

Clusters are represented as numbered points. The diagonal line represents 100 % of discontinuous concatenations. The dashed lines represent 75 % and 25 % of discontinuous concatenations, respectively. Cluster diameter values are in Hz.

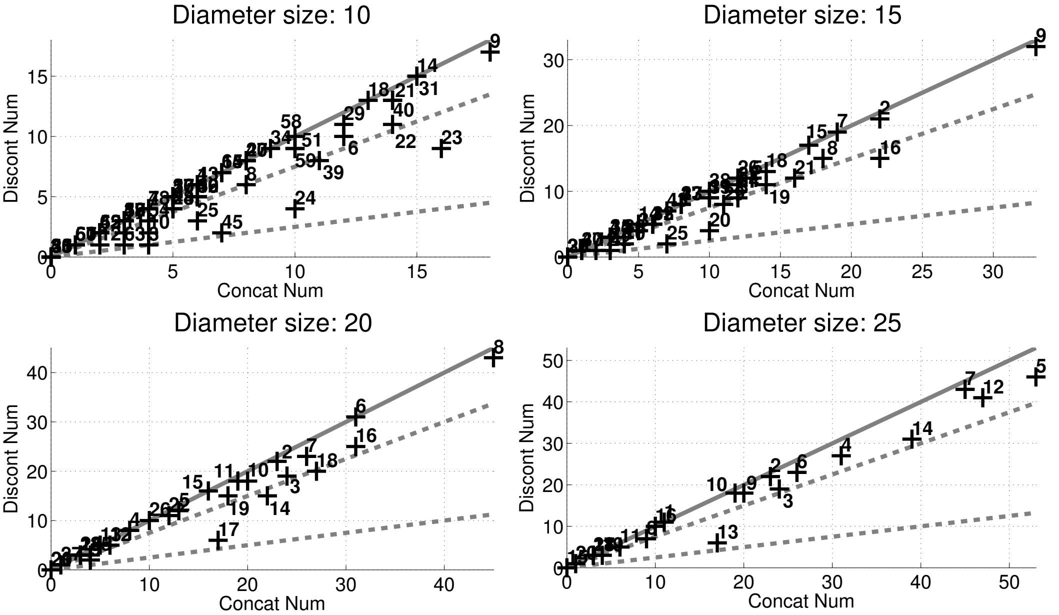


Fig. 3.10: Discontinuous concatenations per cluster - vowel /e/, female speaker

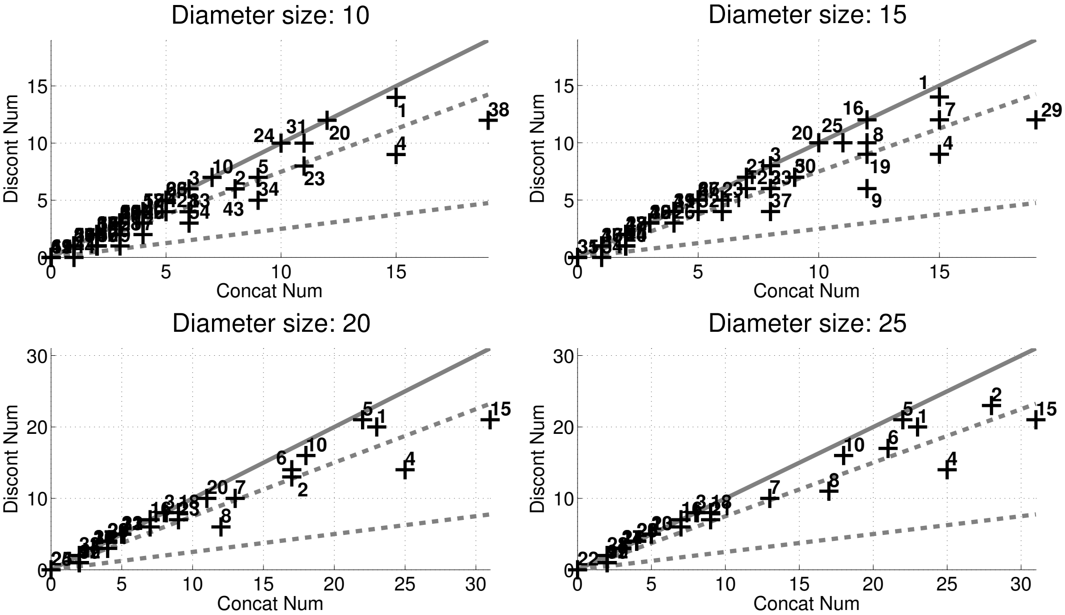


Fig. 3.11: Discontinuous concatenations per cluster - vowel /i/, female speaker

Clusters are represented as numbered points. The diagonal line represents 100 % of discontinuous concatenations. The dashed lines represent 75 % and 25 % of discontinuous concatenations, respectively. Cluster diameter values are in Hz.

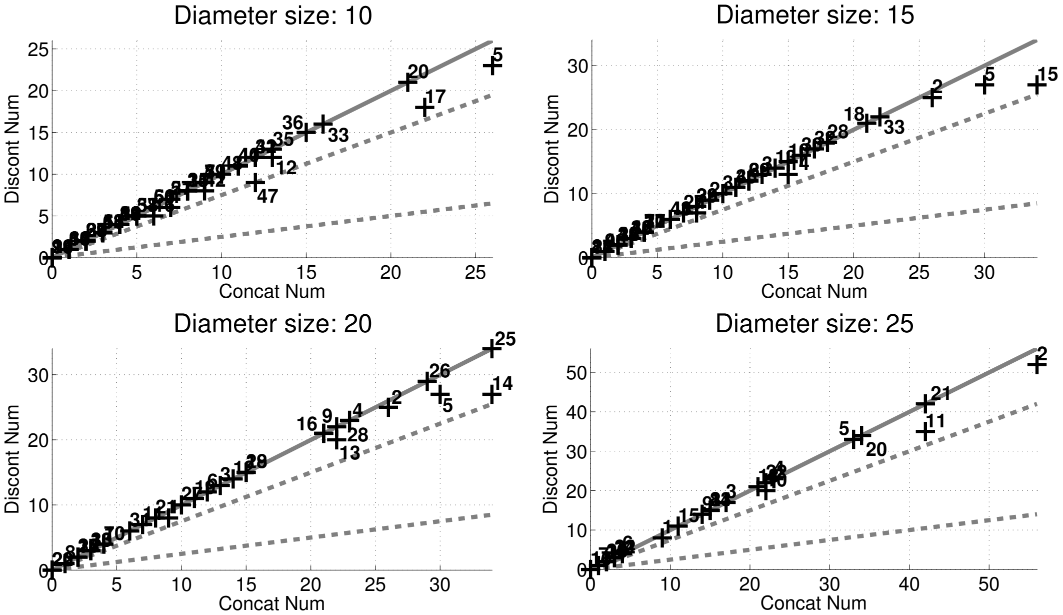


Fig. 3.12: Discontinuous concatenations per cluster - vowel /o/, female speaker

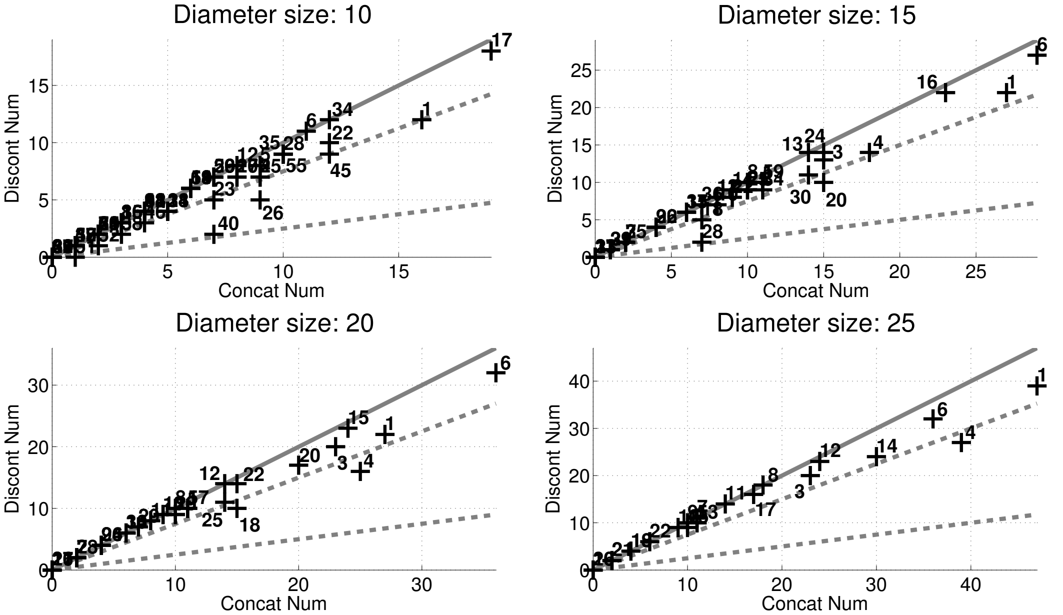


Fig. 3.13: Discontinuous concatenations per cluster - vowel /u/, female speaker

Clusters are represented as numbered points. The diagonal line represents 100 % of discontinuous concatenations. The dashed lines represent 75 % and 25 % of discontinuous concatenations, respectively. Cluster diameter values are in Hz.

contexts potentially present in any cluster may affect our decision regarding the *criticality* of that particular cluster. This is especially important when the halves of sentences are repeatedly used in the concatenations. This aspect should help us to better understand the consonantal contextual phonetic effects or other effects not related to $F0$, supposing that they exist. On the other hand, the larger diameter values lead to more sentences falling into an overall lower number of clusters, which may to some extend make the position of the clusters, with respect to their central $F0$ contour, be the main factor leading to their *criticality*.

Procedure

For the purpose of understanding which of the above mentioned reasons is the more dominant one, basic statistics about the concatenations of the observations contained in *significant clusters* were collected. The *significant clusters* were defined as clusters that are either *critical* or the closest to the *criticality borders* if no *critical clusters* exist for a particular vowel-speaker combination (e.g. /a/, female speaker). The *continuous significant clusters* will be hereafter denoted as `ConS` (`Con`tinuous `S`ignificant) and the *discontinuous significant clusters* as `DisS` (`Dis`continuous `S`ignificant).

For each concatenation, the observations contained in these clusters were involved in, the concatenation cost, defined as the Euclidean distance between the $F0$ contours of the concatenated units, was calculated. Note that this was the same metric as used for calculating dissimilarities between observations during clustering. As a matter of fact, the observations of most of the *significant clusters* are involved in both continuous and discontinuous concatenations, and as such can be further divided into two groups based on *continuity* for each *criticality* type, which means that four groups can be obtained in total. In the analysis, we look separately at the *discontinuous concatenations* of the observations contained in the *discontinuous significant clusters* (`DisS_D`), the *continuous concatenations* of the observations of the *discontinuous signif-*

icant clusters (DisS_C), the *discontinuous concatenations* of the observations contained in the *continuous significant clusters* (ConS_D), and the *continuous concatenations* of the items of the *continuous significant clusters* (ConS_C). Note that within cluster concatenations were not taken into account in this analysis.

Having all concatenations divided into the groups, a maximum concatenation cost, a minimum concatenation cost, a median concatenation cost, a mean concatenation cost and a standard deviation of concatenation costs of each group were calculated.

The mean value of the concatenation costs for a vowel v, a cluster *criticality* $x \in \{\mathtt{DisS}, \mathtt{ConS}\}$ and a concatenation *continuity* $y \in \{\mathtt{C}, \mathtt{D}\}$ can formally be described as:

$$\widehat{\mu}_{S^v_{x\_y}} = \frac{1}{N} \sum_{S^v \in S^v_x} \sum_{\substack{O^v_i \in S^v \\ O^v_j \notin S^v}} D\left(O^v_i, O^v_j\right) \Big| C\left(O^v_i, O^v_j\right) \in facts_y, \qquad (3.5)$$

where N is a number of observations, S^v_x is a set of all *significant clusters* of the vowel v and the *criticality* x, D stands for the Euclidean distance, O^v denotes individual observations of the vowel v, C stands for a concatenation operation and $facts_y$ is a subset of *facts* (as defined in Sec. 2.4.4) of the *continuity* y. The other statistics can be defined analogically.

Results

If the eccentricity of the *significant clusters* was the main reason for their significance, the mean concatenation costs of the *discontinuous significant clusters* would be larger than those of the *continuous significant clusters.* Also, if the distances between the observations were more important than the possible phonetic or other effects, the means of the concatenation costs of the continuous and the discontinuous concatenations would differ, no matter what type

Tab. 3.6: Euclidean distances of $F0$ contours - *significant clusters*, male speaker

		DisS_D	DisS_C	ConS_D	ConS_C
/a/	max	56.93	NA	41.48	28.18
	min	8.69	NA	11.52	9.31
	median	50.67	NA	24.16	13.61
	mean	42.94	NA	24.83	16.30
	std	20.16	NA	10.18	5.80
/e/	max	67.78	31.7	63.78	38.9
	min	7.94	8.88	13.27	8.02
	median	26.49	14.05	39.32	13.27
	mean	27.20	18.31	40.00	15.43
	std	16.30	9.35	22.95	7.48
/i/	max	85.58	36.16	42.02	34.68
	min	11.32	12.99	11.32	10.15
	median	30.30	25.17	25.95	19.36
	mean	35.72	25.37	25.67	19.53
	std	23.51	9.46	10.17	8.73
/o/	max	84.45	13.55	69.56	39.77
	min	12.71	12.71	13.55	9.76
	median	40.01	13.13	28.41	13.55
	mean	38.10	13.13	36.40	18.98
	std	20.96	0.59	16.31	9.59
/u/	max	70.31	24.91	25.31	25.31
	min	9.85	12.22	20.8	11.78
	median	35.67	15.86	23.06	23.06
	mean	36.85	17.71	23.06	20.8
	std	18.60	5.61	3.19	6.38

of *criticality* we look at (cf. Dis_C vs. Dis_D and Con_C vs. Con_D in Tab. 3.6 and Tab. 3.8).

The results shown in Tab. 3.6 and Tab. 3.8 suggest that for some of the cases the distances of the $F0$ contours indeed seem to be the dominant factor. To formally verify this observation, the means of the concatenation costs were compared using the two sample Student's t-test. Since the variances of the concatenation costs of different groups were not equal, they were estimated separately, and the Welch approximation to the degrees of freedom was used

Tab. 3.7: Statistical comparison of means ($\alpha = 0.05$) - *significant clusters*, male speaker

/a/	DisS_D vs. DisS_C	NA
	DisS_D vs. ConS_D	$t = 1.82$, $df = 5.68$, $p = 0.0604$
	DisS_D vs. ConS_C	$t = 2.91$, $df = 4.22$, $p = 0.0204$*
	DisS_C vs. ConS_C	NA
	ConS_D vs. ConS_C	$t = 1.93$, $df = 6.34$, $p = 0.0496$*
/e/	DisS_D vs. DisS_C	$t = 1.79$, $df = 13.17$, $p = 0.0486$*
	DisS_D vs. ConS_D	$t = -1.29$, $df = 6.22$, $p = 0.1211$
	DisS_D vs. ConS_C	$t = 3.18$, $df = 37.68$, $p = 0.0015$*
	DisS_C vs. ConS_C	$t = 0.68$, $df = 7.54$, $p = 0.2588$
	ConS_D vs. ConS_C	$t = 2.57$, $df = 5.40$, $p = 0.0232$*
/i/	DisS_D vs. DisS_C	$t = 1.90$, $df = 17.59$, $p = 0.0373$*
	DisS_D vs. ConS_D	$t = 1.77$, $df = 15.77$, $p = 0.0478+$
	DisS_D vs. ConS_C	$t = 3.59$, $df = 49.00$, $p = 0.0004$*
	DisS_C vs. ConS_C	$t = 1.30$, $df = 8.86$, $p = 0.1140$
	ConS_D vs. ConS_C	$t = 1.29$, $df = 8.33$, $p = 0.1160$-
/o/	DisS_D vs. DisS_C	$t = 8.58$, $df = 52.99$, $p = 0$*
	DisS_D vs. ConS_D	$t = 0.39$, $df = 59.35$, $p = 0.3485$
	DisS_D vs. ConS_C	$t = 4.69$, $df = 33.37$, $p = 0$*
	DisS_C vs. ConS_C	$t = -2.00$, $df = 10.38$, $p = 0.0361+$
	ConS_D vs. ConS_C	$t = 4.00$, $df = 30.84$, $p = 0.0002$*
/u/	DisS_D vs. DisS_C	$t = 6.66$, $df = 19.33$, $p = 0$*
	DisS_D vs. ConS_D	$t = 4.64$, $df = 3.01$, $p = 0.0094+$
	DisS_D vs. ConS_C	$t = 4.30$, $df = 5.61$, $p = 0.0030$*
	DisS_C vs. ConS_C	$t = -0.81$, $df = 5.69$, $p = 0.2262$
	ConS_D vs. ConS_C	$t = 0.58$, $df = 3.86$, $p = 0.2979$-

where necessary. The results of the statistical test are shown in Tab. 3.7 and Tab. 3.9[5].

If our hypothesis of the eccentricity of the clusters and thus the importance of $F0$ is valid, statistically significant differences should exist between sets DisS_D vs. DisS_C, ConS_D vs. ConS_C and DisS_D vs. ConS_C. In other words,

[5] The '*' symbol denotes lines supporting the hypothesis, the '-' symbol denotes differences disconfirming the hypothesis and '+' denotes differences that do not necessarily disconfirm the hypothesis.

Tab. 3.8: Euclidean distances of $F0$ contours - *significant clusters*, female speaker

		`DisS_D`	`DisS_C`	`ConS_D`	`ConS_C`
/a/	max	133.58	34.14	82.84	37.01
	min	17.85	34.14	12.47	15.33
	median	64.68	34.14	41.68	33.19
	mean	62.14	34.14	45.04	28.51
	std	32.59	NA	27.30	11.57
/e/	max	128.32	30.88	43.78	43.78
	min	11.8	18.47	20.77	16.43
	median	39.72	20.13	28.31	20.78
	mean	46.41	22.40	29.63	24.61
	std	30.25	5.71	7.90	10.11
/i/	max	168.99	55.07	124.13	47.13
	min	13.97	55.07	13.09	16.84
	median	95.99	55.07	64.90	32.17
	mean	83.92	55.07	63.00	29.03
	std	46.15	NA	39.88	12.69
/o/	max	187.58	NA	140.90	27.21
	min	14.51	NA	14.51	27.21
	median	45.27	NA	31.63	27.21
	mean	63.72	NA	41.92	27.21
	std	48.04	NA	32.48	NA
/u/	max	172.09	14.65	120.00	16.16
	min	13.82	14.65	14.69	14.69
	median	38.44	14.65	29.59	15.43
	mean	52.37	14.65	44.87	15.43
	std	39.48	NA	33.47	1.04

we would like to reject the null hypothesis that there is zero difference in means of these groups. Differences in means between sets `DisS_D` vs. `ConS_D` and `DisS_C` vs. `ConS_C` do not necessarily disconfirm the hypothesis.

Let us first look at the results of the statistical analysis obtained for the male voice. These results are summarized in Tab. 3.7. It can be seen that for the vowels /a/, /e/ and /o/, the $F0$ differences at the concatenation points as measured by the defined concatenation cost are the dominant factor. The statistically significant difference obtained for the vowel /o/ between groups

Tab. 3.9: Statistical comparison of means ($\alpha = 0.05$) - *significant clusters*, female speaker

/a/	DisS_D vs. DisS_C	NA
	DisS_D vs. ConS_D	$t = 1.64$, $df = 9.52$, $p = 0.0668$
	DisS_D vs. ConS_C	$t = 4.33$, $df = 3.63$, $p = 0.0076$*
	DisS_C vs. ConS_C	NA
	ConS_D vs. ConS_C	$t = 1.41$, $df = 8.49$, $p = 0.0973$-
/e/	DisS_D vs. DisS_C	$t = 5.02$, $df = 20.4$, $p = 0$*
	DisS_D vs. ConS_D	$t = 1.93$, $df = 6.34$, $p = 0.0496$+
	DisS_D vs. ConS_C	$t = 3.34$, $df = 25.11$, $p = 0.0013$*
	DisS_C vs. ConS_C	$t = -0.46$, $df = 8.98$, $p = 0.3274$
	ConS_D vs. ConS_C	$t = 1.00$, $df = 10.93$, $p = 0.1683$-
/i/	DisS_D vs. DisS_C	NA
	DisS_D vs. ConS_D	$t = 1.39$, $df = 17.48$, $p = 0.0916$
	DisS_D vs. ConS_C	$t = 5.46$, $df = 24.43$, $p = 0$*
	DisS_C vs. ConS_C	NA
	ConS_D vs. ConS_C	$t = 2.46$, $df = 11.92$, $p = 0.0152$*
/o/	DisS_D vs. DisS_C	NA
	DisS_D vs. ConS_D	$t = 2.71$, $df = 73.30$, $p = 0.0042$+
	DisS_D vs. ConS_C	NA
	DisS_C vs. ConS_C	NA
	ConS_D vs. ConS_C	NA
/u/	DisS_D vs. DisS_C	NA
	DisS_D vs. ConS_D	$t = 0.84$, $df = 50.29$, $p = 0.2038$
	DisS_D vs. ConS_C	$t = 6.50$, $df = 48.97$, $p = 0$*
	DisS_C vs. ConS_C	NA
	ConS_D vs. ConS_C	$t = 4.20$, $df = 22.43$, $p = 0.0002$*

DisS_C vs. ConS_C is actually due to lower mean value of the DisS_C, which is still in favor of our hypothesis. For the vowels /i/ and /u/, the null hypothesis cannot unfortunately be rejected, which means that there may exist other important factors influencing human judgments than the distances of the $F0$ contours. Another observation that can be made is that, despite the $F0$ main role proved by the statistical analysis, the min values of the *discontinuous concatenations* can be very low, which again confirms that other effects also exist for the vowels /a/, /e/ and /o/, but rather particularly. More can potentially be revealed by the analysis of phonetic contextual effects presented in the next chapter.

If we turn next to the female voice results listed in Tab. 3.9, very interesting observation can be made. The obtained results are in contrast to those for the male voice. For the vowels /a/ and /e/, there seem to be other important factor(s) than the $F0$ differences as the null hypothesis cannot be rejected for the groups `ConS_D` vs. `ConS_C`. For the vowels /i/ and /u/, we do not unfortunately have observations in `Dis_C` set. Therefore, not as clear conclusion can be made as for the male voice vowels /a/, /e/ and /o/. It can however be seen that for the groups `ConS_D` vs. `ConS_C`, the null hypothesis can be rejected, which was not the case for the male voice.

The similarity of the two voices can be found in the low minimum values for the *discontinuous concatenations*, which suggests that for both voices discontinuities not necessarily related only to $F0$ can be found[6].

For the female vowel /o/, no results could be obtained by the analysis due to low number of continuous concatenations in the test data.

Discontinuity and Continuity Thresholds

Once the role of $F0$ was analyzed by comparing the means of the concatenation costs from the *criticality* and the *continuity* point of view for each particular vowel-speaker combination, it became interesting to do a similar analysis comparing individual vowels and to look for synergies for the two speakers. The question was if there exist for different vowels the same thresholds that would allow for separating likely continuous from likely discontinuous concatenations. And, in case that the thresholds differ from vowel to vowel, whether or not some similarities can be found when comparing the two speakers. To answer these questions, statistical comparisons of the means of the concatenation costs were again made, but this time, the groups `DisS_D` and `ConS_C` were taken into

[6] At least in the sense of the defined concatenation cost function.

Tab. 3.10: Statistical comparison of means across vowels ($\alpha = 0.05$)

DisS_D vs. DisS_D	male	/a/ vs. /e/	$t = 1.64$, $df = 5.06$, $p = 0.0802$
		/a/ vs. /i/	$t = 0.74$, $df = 5.58$, $p = 0.2458$
		/a/ vs. /o/	$t = 4.20$, $df = 22.43$, $p = 0.0002$*
		/a/ vs. /u/	$t = 0.51$, $df = 4.85$, $p = 0.3159$
		/e/ vs. /i/	$t = -1.70$, $df = 61.00$, $p = 0.0473$*
		/e/ vs. /o/	$t = -2.53$, $df = 62.31$, $p = 0.0069$*
		/e/ vs. /u/	$t = -2.58$, $df = 45.13$, $p = 0.0066$*
		/i/ vs. /o/	$t = -0.49$, $df = 71.74$, $p = 0.3114$
		/i/ vs. /u/	$t = -0.26$, $df = 55.04$, $p = 0.3977$
		/o/ vs. /u/	$t = 0.36$, $df = 98.35$, $p = 0.3594$
	female	/a/ vs. /e/	$t = 2.85$, $df = 127.96$, $p = 0.0025$*
		/a/ vs. /i/	$t = -2.37$, $df = 44.17$, $p = 0.0111$*
		/a/ vs. /o/	$t = -0.24$, $df = 141.12$, $p = 0.4064$
		/a/ vs. /u/	$t = 1.42$, $df = 91.00$, $p = 0.0794$
		/e/ vs. /i/	$t = -4.11$, $df = 43.29$, $p = 0$*
		/e/ vs. /o/	$t = -2.63$, $df = 136.36$, $p = 0.0047$*
		/e/ vs. /u/	$t = -0.87$, $df = 87.97$, $p = 0.1925$
		/i/ vs. /o/	$t = 2.05$, $df = 56.40$, $p = 0.0225$*
		/i/ vs. /u/	$t = 3.15$, $df = 56.62$, $p = 0.0013$*
		/o/ vs. /u/	$t = 1.46$, $df = 116.44$, $p = 0.0733$
ConS_C vs. ConS_C	male	/a/ vs. /e/	$t = 0.37$, $df = 28.06$, $p = 0.3587$
		/a/ vs. /i/	$t = -1.16$, $df = 22.40$, $p = 0.1288$
		/a/ vs. /o/	$t = -0.82$, $df = 15.30$, $p = 0.2121$
		/a/ vs. /u/	$t = -1.28$, $df = 4.42$, $p = 0.1324$
		/e/ vs. /i/	$t = -1.37$, $df = 25.82$, $p = 0.0912$
		/e/ vs. /o/	$t = -1.03$, $df = 18.01$, $p = 0.1581$
		/e/ vs. /u/	$t = -1.45$, $df = 5.29$, $p = 0.1014$
		/i/ vs. /o/	$t = 0.15$, $df = 20.56$, $p = 0.4421$
		/i/ vs. /u/	$t = -0.32$, $df = 6.63$, $p = 0.3784$
		/o/ vs. /u/	$t = -0.42$, $df = 8.28$, $p = 0.3414$
	female	/a/ vs. /e/	$t = 0.51$, $df = 3.40$, $p = 0.3215$
		/a/ vs. /i/	$t = -0.06$, $df = 4.70$, $p = 0.4776$
		/a/ vs. /o/	NA
		/a/ vs. /u/	$t = 1.95$, $df = 2.05$, $p = 0.0940$
		/e/ vs. /i/	$t = -0.65$, $df = 7.43$, $p = 0.2687$
		/e/ vs. /o/	NA
		/e/ vs. /u/	$t = 2.36$, $df = 6.40$, $p = 0.0269$*
		/i/ vs. /o/	NA
		/i/ vs. /u/	$t = 2.38$, $df = 4.13$, $p = 0.0371$*
		/o/ vs. /u/	NA

account, and the comparisons were made across different vowels. The results obtained by the analysis are presented in Tab. 3.10, in which the statistically significant differences at the significance level $\alpha = 0.05$ are marked by the '*' symbol.

When we look at the results obtained for the male voice, it can be seen that the mean value of the `DisS_D` group for the vowel /e/ is significantly lower than for all other vowels except of the vowel /a/, for which the statistical significance cannot be proven at the given significance level despite the mean value also being higher than for the vowel /e/. Since the mean value of the vowel /a/ is the highest of all, we believe that the difference is at least "practically" significant and could probably be statistically proven having more data on hand. Regarding the `ConS_C` group, no differences were found for the male speaker.

More observations can be made for the female voice. As an interesting result, the vowel /e/ was found to behave similarly to the male speaker. Here, the only difference which cannot be proven statistically is in the pair with the vowel /u/. At this point we however do not want to make any significance claims as the mean value of the vowel /u/ is also comparatively, albeit not significantly, lower than for the remaining vowels. For the `ConS_C`, the vowel /u/ was found to have lower mean value. Statistical significance cannot be shown in the pair with the vowel /o/, which is again due to the low number of continuous observations for this particular vowel, and also in the pair with the vowel /a/, but at this point we can make the same statement about the significance as we have already made for the male voice since again the mean value of the vowel /a/ is approximately twice higher than for the vowel /u/, similarly to all other vowels.

As a clear outcome of this analysis, it can be said that if one introduces continuity and discontinuity thresholds for the two speakers based on the $F0$ differences, the threshold for the vowel /e/ has to be set lower than for other vowels. This statement cannot unfortunately be generalized due to the lim-

Tab. 3.11: Statistical comparison of means for continuous concatenations ($\alpha = 0.05$)—male vs. female speaker

ConS_C	/a/ male vs. /a/ female	$t = -1.78$, $df = 2.21$, $p = 0.1024$
	/e/ male vs. /e/ female	$t = -2.16$, $df = 9.01$, $p = 0.0297$*
vs.	/i/ male vs. /i/ female	$t = -1.55$, $df = 5.42$, $p = 0.0888$
ConS_C	/o/ male vs. /o/ female	NA
	/u/ male vs. /u/ female	$t = 1.64$, $df = 3.30$, $p = 0.0954$

itation of our experiments to two concrete speakers, but it can serve as a hypothesis for future experiments.

Continuity Perception across Speakers

The $F0$ of the female voice is higher than the one of the male voice. Also the dynamics of the $F0$ changes is higher for the female voice. Therefore, one could expect that the continuity threshold when comparing the two voices is higher for the female voice. Tab. 3.6 and Tab. 3.8 show that the expectation is mostly supported by the obtained measurements. The only exception is the vowel /u/ for the male speaker, for which the mean value of the concatenation costs in the group ConS_C is actually higher than for the female voice. For all other vowels, the trend follows our expectation.

In order to verify the statistical significance of this observation, the t-test was again used to compare the mean values of the ConS_C sets for the two speakers. The result of this comparison is presented in Tab. 3.11. Unfortunately, only the observed trend for the vowel /e/ was found to be statistically significant at $\alpha = 0.05$. This keeps us in uncertainty whether the incoherence of the $F0$ contours should be measured absolutely or relatively with respect to the fundamental frequency of individual speakers.

3.4 Conclusions

The role of $F0$ discontinuities in the quality of mid-vowel concatenations has been investigated in this chapter. It has been shown that audible concatenation artifacts can be detected with a high accuracy using SVM classifiers trained on fine-grained $F0$ contours extracted from the concatenation areas of concatenated diphones. The fine-grained $F0$ contours can also be parametrized using linear regression without a considerable loss of the classification accuracy. The SVM classifiers are however difficult to incorporate into our TTS system.

As an alternative, clustering of the $F0$ contours using the Euclidean distance as a metric has been proposed. This technique unfortunately did not prove to be useful. Using perceptually motivated scales—mel and semitone—for representing the contours did not lead to better results.

For some vowels, statistically significant differences have been found between average distances of the $F0$ contours calculated using the Euclidean distance for the continuous and the discontinuous concatenations. This suggests that for those vowels, many of the discontinuities can be predicted by such a simple measure. It is however important to use the $F0$ contours instead of a single point on either side of the concatenation point (as shown in Fig. 2.8). The $F0$ contours allow for capturing the dynamics of $F0$ in contrast to the static differences measured by the single points. The results show that the $F0$ dynamics plays a crucial role.

The main observations from the results presented in this chapter can be summarized as follows:

- Fine-grained $F0$ contours or their approximation are good predictors of mid-vowel audible concatenation discontinuities.
- $F0$ dynamics needs to be measured for both concatenated diphones. It gives better results than static $F0$ differences at the concatenation points.
- Clustering of $F0$ contours using Euclidean distance is not useful for improving the quality of the mid-vowel concatenations.

- For the male voice vowels /a/, /e/, /o/ and the female voice vowels /i/ and /u/, calculating the Euclidean distance of the $F0$ contours of the concatenated units is sufficient for separating most of the continuous and the discontinuous concatenations.
- For the vowel /e/, both voices, the affordable threshold of the $F0$ contour distances for continuous concatenations is lower than for all other vowels.

Chapter 4

Phonetic Analysis

4.1 Introduction

Phonetic features have been reported as being good predictors of concatenation artifacts [KT02]. Generally speaking, phonetic context information is typically embedded as a sub-cost in most of the traditional unit selection implementations. This sub-cost is however often only binary, i.e. agreement/disagreement of a target and a candidate unit contexts. This is mainly due to limited knowledge of what the context mismatch driven latent phenomena that influence human perception are. The goal of the experiments presented in the following sections is to provide more insight into acoustic effects different consonantal contexts can have on vowels. Understanding of these effects is in our opinion crucial for bringing the quality of synthetic speech to a next level, no matter which method is used to build a synthesizer.

4.2 Role of Nasal Contexts

4.2.1 Motivation

[after [LS12]>][1] From the articulatory perspective, nasalized vowels are quite simple—they differ from their oral counterparts only in lowering the velum. However, a large and complex resonance space emerges as a result of opening the nasal cavity, which is why nasalized vowels and vowels in the context of nasals in general represent from the acoustical perspective probably the most complicated sounds of human speech. This simple articulatory gesture is acoustically manifested in various ways, depending especially on the quality of the vowel and also on the degree of acoustic coupling between the two resonance chambers. For these reasons—and also because the nasal cavity and the paranasal cavities of every speaker are different—there are only few universal acoustic correlates of nasality.

The acoustic complex of nasalized vowels consists of nasal formants (formants of the pharyngonasal tract), oral or vocalic formants (whose frequency may, however, be shifted compared to non-nasal vowels), and antiformants which frequently appear in pairs with the nasal formants [FL71].

The most important acoustic features responsible for the perceptual impression of nasality appear in low frequencies. One of the main correlates of nasality, regardless of vowel quality, is the relative lowering of the intensity of F1, specifically by 6–8 dB according to [HS56]. The second "universal" feature related to the nasality is the presence of a spectral peak around 250 Hz. This peak corresponds to the first formant of the pharyngonasal tract and is typically marked as N1. The presence of antiformants due to coupling of the nasal cavity is also universal, but their specific frequencies differ in various studies (see [HS85] for a review).

For the purposes of concatenating units in a TTS system, the presence

[1] Beginning of a part of this thesis borrowed from the co-authored publication [LS12].

of nasality in only one of two concatenated diphones may lead to perceived discontinuities, which was indeed observed in our informal experiments. A nasalized vowel[2] may, on the one hand, manifest higher intensity in low frequencies (around 250 Hz, the N1) and, on the other hand, the energy roll-off above this peak is likely to be considerably greater due to the weaker F1 and generally stronger spectral slope.

Our hypothesis was that it may be undesirable to concatenate a nasalized vowel with a non-nasal vowel or vice versa, since the energy difference in specific frequency bands may cause the impression of discontinuity. The aim was thus to examine whether controlling for the nasality conflict can lead to better continuity of concatenations in vowels. [< after [LS12]][3]

4.2.2 Definition of Nasalization Mismatch

For the analysis, data obtained by the *half-sentence* method described in Sec. 2.2 were used. Since the analysis deals with nasal contexts, a selection was made that only included synthetic sentences containing a *nasalization mismatch*, henceforth referred to as a NAMI set (NAMI stands for the NAsalization MIsmatch). The rest of the sentences formed a NOMI set (NO MIsmatch).

The *nasalization mismatch* was defined as a disagreement between the original context of a synthesized vowel and its target context with respect to the presence of a nasal consonant, no matter if the disagreement was on the left or on the right side of the synthesized vowel. As an example, let us take two words `/t_San/` and `/t_Sas/`, and create a synthetic word `/t_Sa-as/` (dash symbol marks the concatenation point) by combining the first half of the first word with the second half of the latter one. This synthetic word would be considered as containing the *nasalization mismatch*, because the original right

[2] Since the Czech vocalic system do not contain nasalized vowels, we use this term throughout this section to refer to vowels standing in a context of nasals, i.e. either preceded or followed by a nasal consonant.

[3] End of the part of this thesis borrowed from the co-authored publication [LS12].

context of the vowel /a/ in the first word was the nasal /n/, whereas in the synthetic word the right context is the fricative /s/.

4.2.3 Mitigating the Role of F0 Discontinuities

As shown in the previous chapter, F0 discontinuities are unquestionably a significant source of concatenation artifacts. In order to be able to analyze the effect of nasal contexts, the F0 discontinuities have to be factored out. In most related studies, the standard procedure is to smooth the concatenation points with respect to differences in pitch and energy to make sure that any perceived discontinuity is not due to pitch or energy "jumps" at the concatenation points. For the reasons discussed in Sec. 1.8.2, we have decided not to apply any pitch smoothing algorithm during concatenation. Instead, thanks to having a large set of experimental synthetic data available, a selection of sentences that do not contain $F0$ discontinuities can be made.

It has been shown in Sec. 3.3 that clustering of pitch contours and concatenating words whose pitch contours fall within the same clusters is not a reliable way of avoiding $F0$ related concatenation artifacts. Still, the information contained in the pitch contours can be leveraged for predicting concatenation discontinuities with a high accuracy using machine learning techniques as described in Sec. 3.2.

The SVM models trained on fine grained pitch contours extracted from a vicinity of the concatenation points from both concatenated diphones, i.e. the `Nat` set defined in Sec. 3.2.2, were used to identify sentences of the NAMI set that were supposed to be smooth, i.e. not containing an audible discontinuity at the concatenation points. Let us further refer to this set as NAMI-S (NAsalization MIsmatch Smooth). The same models were analogically applied to obtain a NOMI-S set (NO MIsmatch Smooth).

4.2.4 Perceptual Experiment

Test Stimuli

As the next step, a random selection of pairs of synthesized sentences—one sentence from the NAMI-S set and one from the NOMI-S set—containing the same word in the middle was made. Each vowel was represented by 15 pairs of sentences, resulting in the total number of 150 audio samples per voice presented to listeners.

Subjects

The subjects taking part in the listening tests were TTS experts and students working on TTS related projects. There were 9 and 10 subjects who finished the male and the female voice listening tests, respectively. Most of the subjects participated in both listening tests.

Procedure

Two listening tests were organized—one for the male voice and one for the female voice. The listeners were presented with the pairs of audio samples in a randomized order. Their task was twofold: (1) to indicate whether or not they heard a concatenation discontinuity in any of the two samples, (2) to express their preference to one of the two samples in each pair. It was also possible to say that none of the samples was better.

Both listening tests were conducted using a web interface allowing the listeners to work from home. It was, however, stressed in the test instructions that the tests shall be done in a silent environment and using headphones. As a preparation, the participants were presented prior to the listening tests with a couple of samples containing audible discontinuities. There were no restrictions on how many times the listeners played each sentence before providing their answers.

Tab. 4.1: Counts of *discontinuity detection facts* - `DISC_FACT`.

	Female		**Male**	
	NAMI-S	**NOMI-S**	**NAMI-S**	**NOMI-S**
/a/	14	0	0	0
/e/	15	1	1	0
/i/	6	0	0	0
/o/	9	0	2	0
/u/	1	1	0	0

4.2.5 Listening Test Evaluation

Since the reliability of results obtained in any listening test—no matter if the participants are experts or not—is always an issue, a per listener ratings analysis was conducted in line with the procedures described in Sec. 1.8.3. Based on the analysis results, one listener was excluded from each listening test.

As the next step, two sets of *facts* have been collected. The first set contained *discontinuity detection facts*. These were the audio samples for which more than or equal to 80% of listeners indicated that they heard a discontinuity (henceforth refferred to as `DISC_FACT`). The second set—*preference facts*—was based on the preference scores. This set was composed of sentences for which more than or equal to 80% of listeners expressed their preference to one of the samples in a pair or indicated no preference (henceforth referred to as `PREF_FACT`).

The results of the *facts* collection are summarized in Tab. 4.1-4.2. It is obvious from both tables that the *nasalization mismatches* at the concatenation points do not matter for the male voice. This is in contrast to the female voice where we can see that especially for the vowels /a/, /e/ and /o/, there is an impact of the *nasalization mismatch* on the perceived quality of the concatenations. To less extend, the effect can also be found for the vowel /i/.

Tab. 4.2: Counts of *preference facts* - PREF_FACT. None stands for *no preference facts*.

	Female			Male		
	NAMI-S	**NOMI-S**	**None**	**NAMI-S**	**NOMI-S**	**None**
/a/	0	15	0	0	0	3
/e/	0	11	0	0	1	4
/i/	1	5	0	0	1	4
/o/	0	10	0	0	2	6
/u/	2	3	0	0	0	6

4.2.6 Analysis of Results

Discussion

Seeing the results, it was interesting to speculate about the reasons for the obtained values from the perspective of the theory of speech production. Since the problems were observed mainly in the female voice vowels /a/, /e/ and /o/, one interesting hypothesis that arose, was that the perceptual effects were due to a complex interaction of spectral peaks in a low frequency range.

[after [LS12]>][4] For the female speakers' high vowels (/i/ and /u/), there are three spectral peaks in the frequency range between 200 and 500 Hz—fundamental frequency $F0$, as well as the first oral and nasal formants ($F1$ and $N1$, respectively); see, for instance, [HS56]. It is well known that frequency components lying within 3 to 3.5 Bark from each other tend to be perceptually integrated into one broader peak [Chi85]. That is exactly the case with the spectral peaks mentioned above.

It is therefore possible that concatenating an oral vowel and a vowel from a nasal context can result in some sort of discontinuity acoustically, but since it is the energy within this 3.5-Bark band, which is relevant perceptually, the discontinuity can be inaudible. Supposing that this was true, it would explain why /i/ and /u/ behave differently from the other vowels. $F1$ of the other

[4] Beginning of a part of this thesis borrowed from the co-authored publication [LS12].

vowels lies in higher frequencies and therefore falls outside of the 3.5-Bark range of the perceptual integration. This would also explain why a similar situation was not found for the male voice. His $F0$ lies much lower than the peak complex of $F1$ and $N1$. [< after [LS12]][5]

Analysis of Discontinuities

To verify the hypothesis formulated in the previous paragraph, we more closely investigated the concatenation areas in both time and frequency domains, and an intriguing finding was discovered. As shown in Fig. 4.1, the reason for the perceived discontinuities was a phase mismatch at the concatenation points. The phase mismatch appeared as a consequence of misplacement of pitch marks by our pitch marking algorithm [LMT11]. The algorithm gets confused by the strengthening of a harmonic signal component the peaks of which are in a close vicinity of the $F0$ peaks. This strengthening of the harmonic component appears when a vowel (/a/, /e/ or /o/, to be more precise) stands in a context of a nasal consonant. For the vowel /a/, all audible discontinuities can be fixed by manual re-labeling of pitch marks in the concatenation areas as shown in the figure.

The situation is however more complicated for the vowels /e/ and /o/, for which the perceived discontinuities cannot be removed in this way in all sentences. The reason is that the harmonic component mentioned above gets not only stronger, but sometimes also interferes with the $F0$ peaks. An example of such interference is shown in Fig. 4.2. In those cases, it seems to be more advisable to completely avoid concatenations. Interestingly enough, for the male voice this phenomenon was not observed. It explains why the *nasalization mismatch* does not matter for this particular voice.

Regarding the high vowels of the female voice, the discussed harmonic component is very weak in both nasal and non-nasal contexts. More important

[5] End of the part of this thesis borrowed from the co-authored publication [LS12].

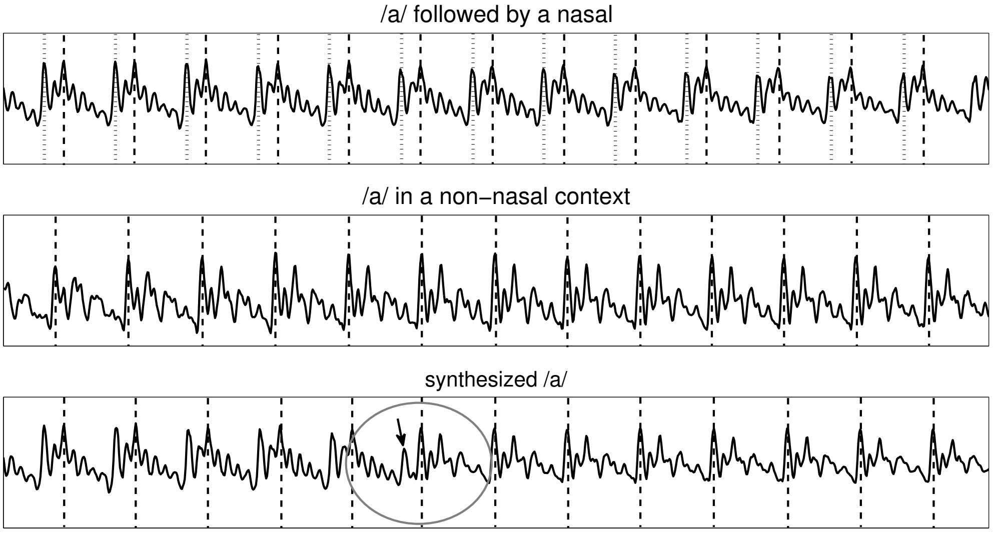


Fig. 4.1: Illustration of a phase mismatch at a concatenation point due to mislabeling of pitch marks

The dashed lines show positions of pitch marks as originally given by the automatic pitch marking algorithm [LMT11]. The dotted lines show manual corrections of the original pitch mark positions. The circled area contains a pointer to an artificial signal peak due to the phase mismatch at the concatenation point.

factor seems to be energy differences at the concatenation points (especially for /i/). It however appears to be perceptually of less importance, as the listeners agreed on the discontinuity *facts* for a smaller number of sentences.

4.2.7 Conclusions

It has been shown that the *nasalization mismatches* only play an important role in the perception of the quality of the mid-vowel concatenations for the female voice vowels /a/, /e/, /o/, and to less extend also /i/. This result nicely supplements the observation given in Sec. 3.4 where we concluded that for the female voice vowels /a/, /e/ and /o/, other $F0$ non-related impacts on the quality of the mid-vowel concatenations may exist.

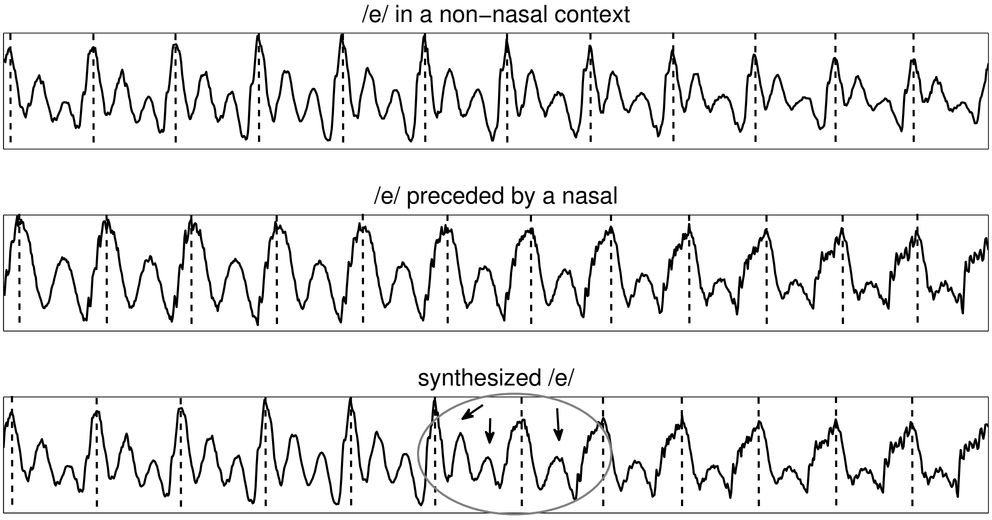


Fig. 4.2: Concatenation discontinuity due to an interference of harmonic components

> *The dashed lines show positions of pitch marks given by the automatic pitch marking algorithm [LMT11]. The concatenation area is depicted by the circle, the arrows are pointing to the "harmonic" mismatch.*

A closer analysis of the discontinuities has shown that they are caused by phase mismatches at concatenation points. The phase mismatches are due to mislabeling of pitch marks and can be manually corrected. It is however difficult to automate this correction. In addition, the phase mismatches are sometimes accompanied with the interference of harmonic components. In those cases, re-labeling of pitch marks is not possible at all.

Therefore, avoiding concatenations of nasalized and non-nasalized vowels /a/, /e/ and /o/ is the simplest way to prevent some concatenation artifacts for the female voice. For the male voice, no special attention needs to be paid to the nasal contexts.

4.3 Possible Impacts of other Phonetic Context Mismatches

4.3.1 Motivation

Since the expert knowledge driven analysis of the *nasalization mismatches* revealed interesting findings, it seemed to be beneficial to more closely investigate other possibly important phonetic context mismatches. For the purposes of this analysis, the perceptual data collected in the large scale listening tests described in Sec. 2.6 was used. This data had not been designed for the analysis of impacts of the phonetic contexts in the first place, but they could still be leveraged for conducting a preliminary analysis of potentially existing phonetic impacts. This was the goal of the experiment described in the following subsections.

4.3.2 Definition of Consonantal Context Mismatch

Similarly to the *nasalization mismatch*, the *consonantal context mismatch* was defined as a disagreement between the original context of a vowel and its target context with respect to the presence of a consonant from a distinct consonantal group, no matter if the disagreement was at the left or at the right context of a synthesized vowel. To give again an example, let us take two words `/t_Sat_s/` and `/t_Sas/`, and create a synthetic word `/t_Sa-as/` (dash symbol marks the concatenation point) using the left part of the first one and the right part of the second one. In the analysis, the synthetic word would be considered as containing twofold phonetic context mismatch—*affricates* and *fricatives.*

4.3.3 Procedure

In order to avoid data scarcity problem, the analysis aimed at investigation of two aspects—the role of a manner of articulation and the role of a place of articulation. All consonants were divided into groups according to their manner and place of articulation, and for each group, a concatenation detection rate was obtained from the set of the *facts* collected in the large scale listening tests.

Note that the limited size of the dataset with respect to different consonantal contexts did not unfortunately allow for mitigating the role of $F0$ discontinuities similarly to how it was done in the *nasalization mismatch* analysis presented in Sec. 4.2. The $F0$ related impacts therefore represent a noise in the data. It was however assumed that this noise is equally distributed across all consonantal contexts. We will further get to this point in Sec. 4.4.5.

4.3.4 Results

Role of Manner of Articulation

In Tab. 4.3, the outcome of the analysis of the impacts of the manner of articulation for both voices is summarized.

The results have been checked for statistical significance[6] using the significance level $\alpha = 0.1$. The χ^2 test for independence has shown that the *nasals mismatches* are a very important factor for the female voice vowels /a/, /e/, and in comparison with the *affricates mismatches* also for the vowel /u/. The *approximants mismatches* have been found to be not as important as the mismatches in *trills* and *plosives* for the female vowel /i/.

Regarding the male voice, the *affricates* and the *approximants mismatches* have shown to lead to significantly smaller amount of discontinuities for the vowels /e/ and /i/, and the *affricates mismatches* also for the vowel /o/. For the same vowel, the *approximants* have however been found to be a very sensitive context.

[6] Note that due to data scarcity, we decided to conduct the statistical analysis at lower confidence level than traditionally used.

Tab. 4.3: Concatenation detection—manner of articulation. Bold numbers show statistically significant observations ($\alpha = 0.1$). The first column for each voice contains a percentage of *discontinuous facts*, the second column a total number of sentences containing a given context mismatch.

Vowel	Phonetic Class	Female		Male	
		[%]	N	[%]	N
/a/	plosives	90.0	110	44.1	34
	nasals	**100.0**	**89**	33.3	15
	fricatives	89.1	137	39.6	48
	affricates	89.9	79	52.8	36
	trills	95.7	23	33.3	9
	approximants	91.7	36	41.7	12
/e/	plosives	88.1	143	57.6	66
	nasals	**100.0**	**68**	53.9	13
	fricatives	85.3	129	61.7	60
	affricates	85.0	20	**25.0**	**16**
	trills	80.0	20	63.6	11
	approximants	88.9	72	**24.1**	**29**
/i/	plosives	**88.1**	**84**	69.4	62
	nasals	86.0	57	54.6	33
	fricatives	84.9	99	72.9	70
	affricates	85.7	21	**11.8**	**17**
	trills	**90.9**	**33**	42.9	7
	approximants	**73.1**	**26**	**18.2**	**11**
/o/	plosives	100.0	116	79.7	69
	nasals	100.0	107	86.1	36
	fricatives	99.3	138	80.6	72
	affricates	100.0	34	**58.3**	**12**
	trills	**96.9**	**32**	66.7	15
	approximants	100.0	35	**100.0**	**15**
/u/	plosives	92.7	124	87.7	106
	nasals	**100.0**	**25**	88.0	25
	fricatives	92.0	113	88.9	90
	affricates	**85.7**	**28**	76.9	13
	trills	**100.0**	**21**	90.0	20
	approximants	93.9	33	95.2	21

Role of Place of Articulation

The results of the analysis of the role of place of articulation are presented in Tab. 4.4.

For the female voice vowel /a/, significantly lower number of discontinuities have been detected for the *labiodentals mismatches* compared to the *prealveolars* and the *palatals mismatches*. The *glottals* have shown to be the least sensitive context for the vowel /e/, whereas the *palatals mismatches* have given rather large number of discontinuities, albeit not statistically significant except of the comparison with the *postalveolars*. For the vowel /i/, the *glottals mismatches* have been found to lead to better concatenations than the *postalveolars* and the *palatals mismatches*. The vowel /u/ has shown to be sensitive to the *bilabials mismatches*, in contrast to the *postalveolars mismatches*, for which a statistical difference from all other groups except of the *labiodentals* and the *glottals*, exists.

We turn next to the results obtained for the male voice. It can be seen that the *glottals mismatches* have lead to larger amount of discontinuities for the vowels /a/, /i/ and /u/. Although, for the vowel /u/, a statistically significant difference has not been found. For the vowel /o/, the impact of the *glottals mismatches* is rather weak, even though a statistical significance has not been found either due to a low number of observations. For the vowel /i/, the *palatals mismatches* are also an important source of discontinuities similarly to the vowel /o/ but in contrast to the vowel /u/.

Voice and Vowel Similarities

It is obvious from the previous subsections that there are many different observations, which are difficult to grasp. In order to find out whether or not there are any similarities between the two speakers and also for different vowels, a correlation analysis of the discontinuity detection rates was conducted.

Tab. 4.4: Concatenation detection—place of articulation. Bold numbers show statistically significant observations ($\alpha = 0.1$). The first column for each voice contains a percentage of *discontinuous facts*, the second column a total number of sentences containing a given context mismatch.

Vowel	Phonetic Class	Female		Male	
		[%]	*N*	[%]	*N*
/a/	bilabials	93.5	46	38.5	13
	labiodentals	**81.0**	**21**	61.5	13
	prealveolars	**93.0**	**114**	46.8	47
	postalveolars	89.3	121	40.4	47
	palatals	**94.7**	**57**	50.0	12
	velars	90.6	85	42.5	40
	glottals	90.0	50	**100.0**	**4**
/e/	bilabials	88.5	96	52.5	40
	labiodentals	85.7	21	80.0	5
	prealveolars	89.7	136	54.7	64
	postalveolars	85.9	85	71.2	52
	palatals	**95.2**	**62**	53.1	32
	velars	90.6	96	67.4	43
	glottals	**66.7**	**18**	83.3	6
/i/	bilabials	86.7	60	50.0	40
	labiodentals	85.0	20	69.2	13
	prealveolars	85.0	100	67.6	71
	postalveolars	**92.2**	*64*	61.1	36
	palatals	**91.9**	*62*	**82.5**	**57**
	velars	86.7	60	64.7	34
	glottals	**77.8**	**18**	**80.0**	**10**
/o/	bilabials	100.0	30	90.0	10
	labiodentals	100.0	27	83.3	18
	prealveolars	99.4	160	83.8	80
	postalveolars	100.0	103	86.4	44
	palatals	98.2	54	**95.8**	**48**
	velars	100.0	71	86.1	43
	glottals	100.0	17	71.4	7
/u/	bilabials	**97.1**	**68**	91.4	58
	labiodentals	81.8	11	84.0	25
	prealveolars	90.2	92	91.4	81
	postalveolars	**69.6**	**23**	87.1	31
	palatals	93.6	93	**80.4**	**56**
	velars	89.4	142	90.8	109
	glottals	85.7	14	100.0	6

This analysis has revealed that for the female voice no statistically significant correlation exists with respect to the place of articulation. There are however strong positive correlations between the vowels /a/ and /e/ ($r = 0.89$, $p = 0.019$), /e/ and /u/ ($r = 0.97$, $p = 0.002$), and /a/ and /u/ ($r = 0.79$, $p = 0.061$) with respect to the manner of articulation.

For the male voice, the manner of articulation has shown to positively correlate for the vowels /a/ and /o/ ($r = 0.75$, $p = 0.086$), and negatively correlate for the pair /e/ and /o/ ($r = -0.75$, $p = 0.087$). Strong negative correlation has also been found for the vowels /e/ and /o/ ($r = -0.89$, $p = 0.003$) with respect to the place of articulation.

The comparison of the two speakers has only shown a negative correlation for the vowel /i/ ($r = -0.92$, $p = 0.001$) and a positive correlation for the vowel /u/ ($r = 0.74$, $p = 0.03$) with respect to the place of articulation.

4.3.5 Discussion

Unlike for the *nasalization mismatch*, which was investigated in the Sec. 4.2, in this analysis the impact of the $F0$ discontinuities at the concatenation points could not be factored out due to limited coverage of different consonantal contexts in the test data. This results in comparatively higher discontinuity detection rates, and for the female vowel /o/, it leads to a limited value of the obtained results as a vast majority of the created *facts* were discontinuous, no matter what the consonantal contexts of the vowels were.

We believe that also the inconsistency in the results for the female vowel /u/, for which the role of the *nasals mismatch* was found to be important in this experiment in contrast to the findings reported in Sec. 4.2, could be explained by the presence of the $F0$ discontinuities. This is also supported by the results of the analysis of the $F0$ impacts described in Sec. 3.4 where we concluded that for the female voice vowels /i/ and /u/, the $F0$ discontinuities are the dominant factor.

4.3.6 Conclusions

The analysis has shown that some contextual mismatches have significant effects on the discontinuity detection rates. These differences have however been found not to be very consistent when comparing the two speakers, which suggests that coarticulatory effects could be speaker dependent. This has also been reported in [vCR96]. Also when comparing different vowels, little can be concluded. The manner of articulation tends to reveal more similarities in discontinuity detection rates across the vowels than the place of articulation mismatches.

It would be very useful if the observations made in this analysis could directly be explained by the theory of speech perception/production. The latent phenomena behind the observations are however too complex. Therefore, additional set of experiments described in the next section was designed with the hope to get better understanding of these phenomena.

4.4 Verification Analysis of Phonetic Contexts

4.4.1 Introduction

In the previous section, some consonantal vowel contexts possibly influencing the quality of the mid-vowel concatenations were identified. The presented results can serve as the first indication of what can be expected when certain consonantal context mismatches occur in synthesized sentences. The results cannot unfortunately be directly related to possible coarticulation effects that could be seen in the power spectrum or phase of the synthesized vowels. This comes as a result of using data in which $F0$ impacts could not be factored out.

In the following subsections, results of a better designed experiment aiming at confirming the observed trends are presented. The experiment only addresses phonetic context mismatches that were found to be important in

Tab. 4.5: Scope of the verification of the role of phonetic context mismatches.

Phonetic Class	Female					Male				
	/a/	/e/	/i/	/o/	/u/	/a/	/e/	/i/	/o/	/u/
approximants			X				X	X	X	X
glottals		X	X			X	X	X	X	X
palatals	X	X	X				X	X	X	X
affricates					X		X	X	X	
labiodentals	X				X		X			
postalveolars			X		X					
bilabials	X				X			X	X	
prealveolars	X									
trills	X	X	X	X	X					

the original data. The overview of the scope of the verification analysis is presented in Tab. 4.5.

4.4.2 Data Preparation

The data for this experiment were prepared in the same way as for the previously described *nasalization mismatch* analysis, i.e. synthetic sentences were created using the *half-sentence* method and analyzed by the SVM models trained on fine grained pitch contours. Only sentences identified as continuous by the classification models were used to find pairs of sentences that contained the same word in the middle—one sentence containing a phonetic mismatch, the other not. For each vowel and phonetic class within the scope of the analysis, 15 pairs of sentences were randomly selected and used as test stimuli in a listening test.

4.4.3 Procedure and Evaluation

The selected pairs of sentences were presented to listeners in a randomized order. The task of the listeners was again twofold: (1) to indicate whether or not they heard a concatenation discontinuity in any of the two samples, (2) to express their preference to one of them. It was also possible to say that both samples were qualitywise the same.

In average, 15 non-expert listeners judged each pair of samples. The minimum number of ratings per pair was 13. Upon the test completion, inter-rater reliability was evaluated in line with the procedure described in Sec. 2.6.4. The number of removed listeners varied in the range of 3–5 per test subset[7].

4.4.4 Results

A small number of discontinuities were detected by the listeners in the verification analysis. This is in contrast to the experiment with the original data reported in the previous section. Taking into account the importance of the $F0$ discontinuities (shown in Sec. 3.3.5) and a good performance of the classification models (summarized in Tab. 3.2), this result could have been expected. Nevertheless, knowing already the impact of the *nasalization mismatch* and obtaining statistically significant observations in the original experiment, the results of the current analysis were disappointing. Overall, very little contextual effects on the quality of mid-vowel concatenations exist according to our results summarized in Tab.4.6-4.7.

The presented results only contain phonetic class and vowel combinations for which listeners agreed at least in a few cases on a presence of the concatenation discontinuities, i.e. were able to create *discontinuity detection* or *preference facts*. Due to very low detection rates, only two groups can be considered to give non-accidental results—female voice *trills* for the vowel /i/ and *bilabials* for the vowel /a/. Since bilabials contain the nasal /m/, the obtained result in fact confirms the results of our previous experiment with the *nasalization mismatches*.

[7] The test subset consisted of samples representing one phonetic class.

Tab. 4.6: Counts of *discontinuity detection facts*. None stands for agreement that none of the presented samples contained discontinuity.

			Mismatch	No-mismatch	None
Female	trills	/i/	7	1	1
	bilabials	/a/	6	3	7
		/u/	2	1	4
	prealveolars	/a/	2	2	6
	postalveolars	/i/	1	2	7
Male	palatals	/e/	2	0	6
		/i/	2	1	4
		/o/	2	0	3
		/u/	0	2	10

Tab. 4.7: Counts of *preference facts*. None stands for *no preference facts*.

			Mismatch	No-mismatch	None
Female	trills	/i/	1	7	1
	bilabials	/a/	2	4	6
		/u/	1	2	1
	prealveolars	/a/	1	1	2
	postalveolars	/i/	1	0	7
Male	palatals	/e/	0	2	5
		/i/	0	3	1
		/o/	0	0	5
		/u/	0	0	5

4.4.5 Conclusions and Future Work

The verification analysis of the impact of the phonetic context mismatches on the quality of mid-vowel concatenations presented in this section has shown that no other *consonantal context mismatches* lead to as high discontinuity detection rates as the *nasalization mismatches*. This is in contrast to the findings reported in the previous section where some significant consonantal contexts have been identified. Since the only difference between the two experiments is the mitigation of the role of the $F0$ discontinuities, it is safe to conclude that the $F0$ discontinuities play different roles for different phonetic contexts. In other words, the hypothesis is that different consonantal contexts may allow for different variability of the $F0$ contours which then results in a larger amount of concatenation artifacts in those contexts. This hypothesis can easily be investigated by calculating characteristics of the $F0$ contours of vowels in different consonantal contexts.

One *consonantal context mismatch* that confirmed its importance in the verification experiment is the context of *thrills* for the female voice vowel /i/. An informal analysis of the discontinuities identified by the listeners has shown that the discontinuities can likely be attributed to the energy differences at the concatenation points. It is important to note at this point that similar observation has been made for the same vowel in the analysis of the *nasalization mismatches.* Investigating the possibility of smoothing the concatenation points with respect to the energy differences remains for our future work.

4.5 Conclusions

In this chapter, three analyses of consonantal contextual effects on the quality of mid-vowel concatenations have been presented. The first analysis was driven by expert knowledge and investigated the role of the *nasalization mismatches.* It has confirmed that nasals can represent an important context for the quality of the concatenations. The other two experiments were aimed to analyze possible impacts of other consonantal contexts with respect to the manner and the place of articulation. They have shown that $F0$ related discontinuities play different roles in different consonantal contexts.

Similarly to the analysis of the role of the $F0$ impacts presented in the previous chapter, the contextual analysis did not reveal any reason for discontinuities in the male voice vowels /i/ and /u/. On the other hand, for the female voice vowels /a/, /e/ and /o/, a conclusion can be made based on the results of the analysis.

In short, the main findings of this chapter can be summarized as follows:

- It is undesirable to concatenate diphones from the oral and the nasal contexts for the female voice vowels /a/, /e/ and /o/. For the male voice, the *nasalization mismatches* as well as other *consonantal context mismatches* have not been found to play a role.
- For the female voice, concatenations in the vowel /i/ can lead to discontinuities in the context mismatches of *nasals* and *thrills*. The reason for the discontinuities seem to be energy differences at the concatenation points.
- $F0$ discontinuities play different roles in different consonantal contexts. This may be attributed to different variability of the $F0$ contours in different consonantal contexts.

Chapter 5

Unit Selection Considerations

5.1 Introduction

The unit selection method has seemed to be getting abandoned as a research topic over the last few years. There is no question about the fact that a huge amount of efforts, including this thesis and all related papers, have already been invested in improving the quality of synthetic speech delivered by unit selection based TTS systems since the introduction of the approach [HB96]. The method has been analyzed from almost all possible angles. Many works have dealt with experiments introducing different speech parameterizations and distances, which could be used for measuring the quality of concatenations (see overview in Sec. 1.4); the target cost components [WM98]; pruning of large unit databases; tuning weights of the costs [CC99]; and last but not least optimizing the search routine to lower the computational costs [TKM10], [SKN08], to name some.

Still, we believe that the most important problem related to the unit selection—the "haphazard" presence of audible artifacts—has not been investigated thoroughly enough. Recently, some papers addressing this particular problem have been published [LLW^{+}10], [LLDW11].

Generally speaking, there are three main sources of these quality jumps.

First, the costs that are used while searching for the optimal sequence of units are not always well correlated with human perception. That was actually the motivation for looking more closely at possible solutions for measuring concatenation artifacts in this thesis. Second, any database, no matter how thoroughly it is verified, contains mislabelings at different levels. Third, the traditional implementation of the search algorithm as described in Sec. 1.3.4 allows, as long as the cost of the whole sequence of units is minimum, for selecting units that should locally be avoided according to their assigned costs. This can especially be observed when the unit database is small, but in theory, it is also possible to encounter this problem for a large database system. Little has been invested in analyzing these audible artifacts in more detail and real understanding of the latent constructs that influence human perception.

The goal of this chapter is to examine the basic unit selection algorithm as such and present some considerations that should be taken into account no matter how any of the costs used in a unit selection based system are exactly designed.

The unit selection method in general was briefly described in Sec. 1.3. At this point, more details on the actual implementation used in our TTS system are added [MTR06]. The design of the current system's costs and their weight settings are investigated using a unique method based on analyzing cost outliers appearing in the optimal sequences of synthesized units. The potential of the proposed method for automating the tuning of the unit selection weights as well as for testing different cost functions is presented at the end of this chapter. A modification to the search algorithm that would allow for better leveraging of the information contained in the costs assigned to individual candidate units is also proposed.

5.2 ARTIC TTS system

[after [LTM12] >][1] ARTIC (Artificial Talker in Czech) is a Czech text-to-speech system developed since 1997 [MTR06]. It is a corpus based system, which makes use of a large carefully designed speech corpus annotated at orthographic, phonetic and prosodic levels. Two speech synthesis methods—fixed-inventory synthesis (aka diphone synthesis or single unit instance synthesis) and unit selection synthesis (aka multiple unit instance synthesis)—are currently implemented [MTR06]. Experiments with the HMM-based speech synthesis method have recently been conducted as well [Han10]. Both the fixed-inventory and the unit selection methods currently use diphones as a basic speech unit. The target and the concatenation costs of the unit selection implementation are defined as described in the following subsections. The total cost is then a simple sum of the two costs.

5.2.1 Concatenation Cost Implementation

The concatenation cost consists of three components—difference in energy, difference in $F0$ and the Euclidean distance of 12 MFCC coefficients [Tih05a]. All the values are z-score normalized in order to align their ranges with the maximum being 1. Moreover, the $F0$ sub-cost is only computed when concatenating diphones at voiced ends. In case that voiced segments are to be concatenated with unvoiced or vice versa, the $F0$ sub-cost is set to 1. For concatenations of unvoiced segments, the cost is set to 0. The values of all the three features are calculated pitch-synchronously. The resulting concatenation cost (5.1) is simply the average of the three components:

$$ConC = \frac{CC_{F0} + CC_{En} + CC_{MFCC}}{3} \tag{5.1}$$

[1] Beginning of a part of this thesis borrowed from the co-authored publication [LTM12].

5.2.2 Target Cost Implementation

To compute the target cost, the following strictly symbolic features are evaluated:

- *suitability for prosodic word position* [RM05]. The feature evaluates the difference in position within prosodic word by a non-linearly increasing penalization [TM06]. This allows to avoid the traditional discrete *initial, middle, final* features and models the positions on a continuous basis in a non-linear way.
- *type of prosodeme* [RM05]. This feature uses simple binary match decision.
- *left and right phonetic context.* This feature, often also used as a subcomponent of the concatenation cost, penalizes disagreements in left and right phonetic contexts of a given diphone. Similarly to the prosodemes, this feature is binary with all related disadvantages of it. However, some analyses have recently been undertaken to overcome this limitation [Leg12].

Each feature is weighted by a heuristically set weight. The *type of prosodeme* is the most prominent and the *phonetic context* the least prominent subcomponent. The value of the target cost (5.2) is then given by a weighted average:

$$TgtC = \frac{\sum_{t=1}^{T} F(t)w(t)}{\sum_{t=1}^{T} w(t)}, \tag{5.2}$$

where $F(t)$ is a feature value, T is a number of the features and $w\,(t)$ are the feature weights. [< after [LTM12]][2]

[2] End of the part of this thesis borrowed from the co-authored publication [LTM12].

5.3 Unit Costs Outlier Detection

5.3.1 Motivation

As already mentioned in the introduction, this chapter aims at the audible artifacts haphazardly appearing in the output of the unit selection systems. Let us assume that the unit selection costs correlate reasonably well with human perception. If this was true, most of the selected units of extreme costs should lead to audible artifacts in the TTS output.

In order to see whether or not such units are being selected at all, the box-and-whisker diagrams (boxplots) were used. The boxplots for every single concatenation cost component and also for the costs as such of the units selected in a test set of utterances are shown in Fig. 5.1. It is obvious from the plots that indeed, some units of rather outlying costs tend to appear in the selected sequences of units.

5.3.2 Perceptual Annotation Experiment

Procedure

Having confirmed that some outlying units exist in the optimal sequences, the next step was to investigate whether or not these units coincide with audible artifacts. For this purpose, an annotation experiment using a set of 50 randomly selected synthesized sentences was designed. The task of listeners was to mark segments of these sentences, which they found unnatural or containing any sort of distortion. The shortest segment that could be annotated was a phoneme. As can be seen in Fig. 5.2, most of the participants were actually marking segments of maximum length 3–5 phonemes.

The test was conducted for two systems `Syst1` and `Syst2`[3]. The same

[3] `Syst1` is a product version of the ARTIC system. `Syst2` is an experimental version of the same system. `Syst2` allows for changing the unit selection costs, their weights and also the search mechanism as such. Its current setting slightly differs from the product system.

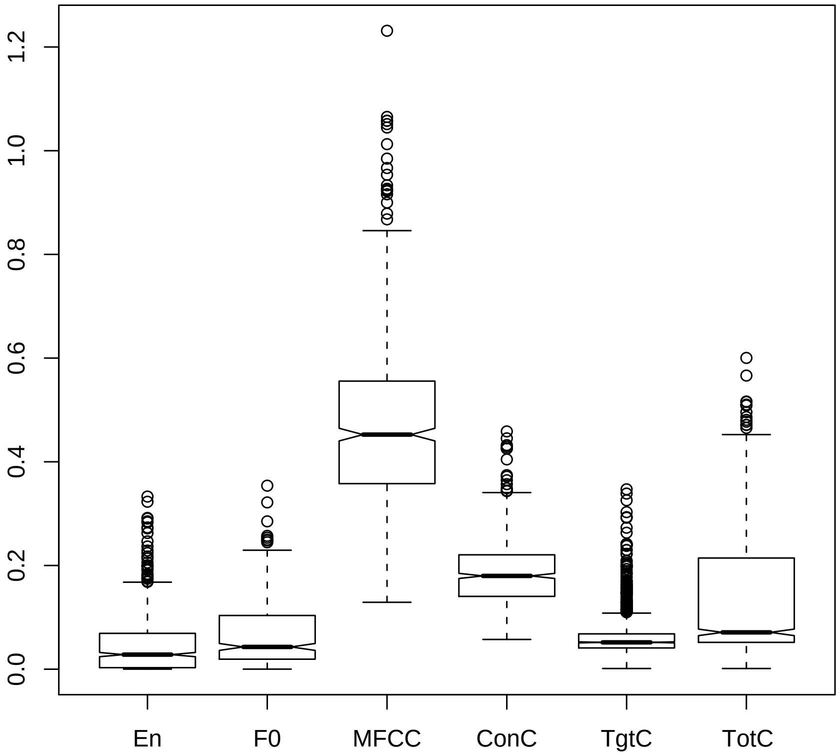


Fig. 5.1: Boxplots of costs of units forming the optimal sequences found by the unit selection search algorithm.

The boxplots show a distribution of the unit selection costs and their sub-components of the selected units in the test set of sentences. The plot whiskers extend out from the boxes to the most extreme data points which are no more than 1.5 times the interquartile range apart from the boxes, hereafter denoted as $1.5 \times IQR$.

listening test web interface as for the experimental data collection described in Chapter 2 was used to conduct the annotation listening tests. It was again stressed in the test instructions that the tests shall be done in a silent environment and using headphones. Since the accurate annotation of audible artifacts is not a simple task, experienced listeners were only invited to participate. In total, 8 listeners, 5 of them being TTS researchers, finished the listening test for the `Syst1`, and 6 listeners (5 TTS experts) annotated the outputs of the `Syst2`.

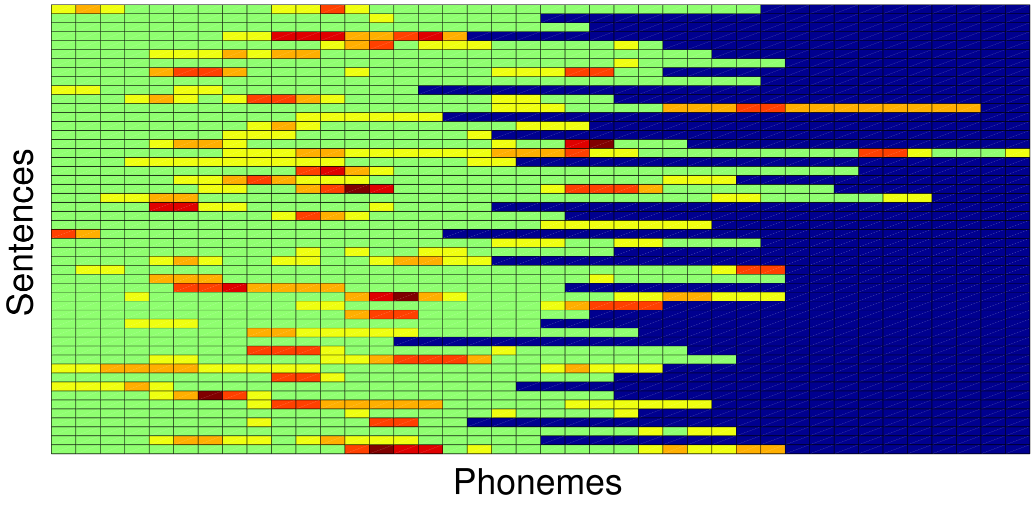


Fig. 5.2: Annotation of artifacts.

The figure shows annotations of artifacts obtained by 6 expert listeners. Each row in the checkerboard plot represents a sentence, each column a phoneme. The darker a particular cell in the plot is the more agreement listeners found regarding a presence of an artifact at the particular phoneme. Note that the test sentences have different lengths, navy blue is used as a figure background.

Listening Test Evaluation

Generally speaking, it is not a simple matter, to evaluate an annotation listening test. One of the typical concerns is how to identify non-reliable listeners. This particular issue was not a problem in our study as all participants were highly motivated to provide good quality annotations.

Another issue is different sensitivity of each participant to various kinds of artifacts. In order to evaluate the perceptual relevance of the outliers, keeping the sensitivity issue in mind, an H_L *score* (5.3) was introduced:

$$H_L\left(i\right) = \frac{\sum_{n=i-L}^{i+L} D_n}{\left(L+1\right)N}. \tag{5.3}$$

L stands for a tolerance interval length in phonemes, i is the index of a phoneme and N is a number of listeners. The number of annotations of the particular phoneme, D_n, is defined as follows:

$$D_n = \sum_{j=1}^{N} h_n\left(j\right). \tag{5.4}$$

$h_n\left(j\right)$ is the annotation of the phoneme n defined as:

$$h_n\left(j\right) = \begin{cases} 1 & n \in A_j \\ 0 & n \notin A_j \end{cases}, \tag{5.5}$$

where the set A_j is the list of indeces of phonemes annotated by the j-th listener.

Having the H_L *score* defined, each outlier can be assigned its value. Since outliers have been defined as units of extreme costs present in the selected sequences of units, i.e. diphones in our case, and the annotations obtained from the listening test are phoneme based, an alignment had to be done. This is reasonable as the concatenation points are located in the middle of phonemes.

Fig. 5.3-5.4 show an illustration of the results of this assignment for the concatenation cost sub-components as well as for the individual costs themselves. Note that based on the above mentioned observation that most of the listeners were using maximum 3–5 phoneme long segments for annotating, the appropriate setting of the maximum length of the tolerance interval can be $L = 2$.

To further quantify the perceptual relevance of the outliers, an ad hoc threshold[4] $thr = 0.5$ was used for the sum of $H_L(i)$ *scores* $S_2(i)$ given by:

$$S_2(i) = \sum_{L=0}^{2} H_L(i). \tag{5.6}$$

[4] More objective setting of the threshold will be presented in Sec. 5.4.1, where additional experiments with this approach are presented.

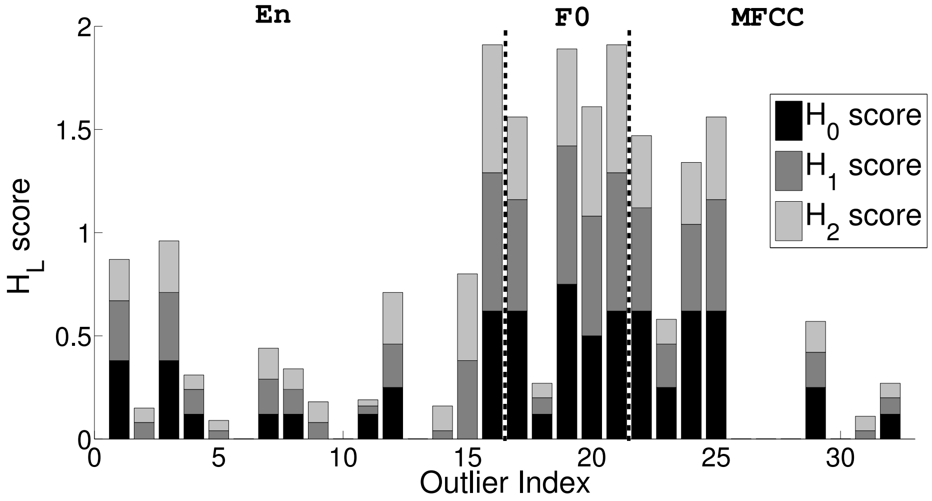


Fig. 5.3: H_L *scores* of outliers of the concatenation cost sub-components (illustrative example).

> *Each detected outlier can be assigned a H_L score using annotations obtained by listeners. The higher the value of the H_L score is the more perceptually relevant the given outlier is. The plot shows outliers detected for the concatenation cost sub-components.*

Using this threshold, each outlier can be assigned a *Hit Rate* value defined as:

$$Hit\ Rate = \frac{N_{hit}}{N_{outl}} \times 100\,[\%]\,, \tag{5.7}$$

where N_{hit} is a number of outliers of a given cost or a cost sub-component for which the condition $S_2(i) \geq thr$ is fulfilled[5], and N_{outl} stands for a number of all outliers found for a given cost or a cost sub-component.

The results of this calculation are summarized in Tab. 5.1-5.2. In the same tables, the percentage of missed *annotated artifacts*—*Missed Rate*—is also presented. The *annotated artifact* has been defined as a phoneme fulfilling

[5] Note that the summation of the scores up to the length L allows for normalizing the relevance of artifacts annotated exactly at a particular phoneme with those annotated less precisely.

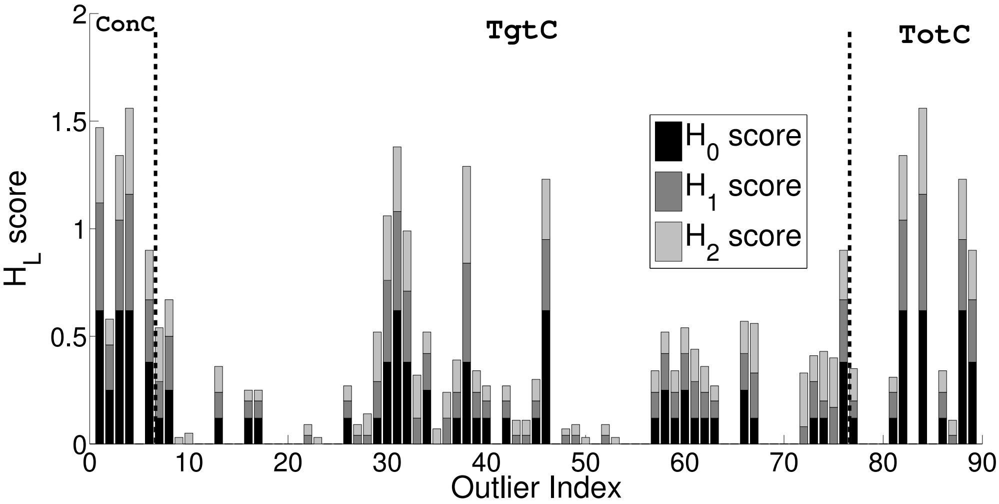


Fig. 5.4: H_L *scores* of outliers of the unit selection costs (illustrative example).

Each detected outlier can be assigned a H_L score using annotations obtained by listeners. The higher the value of the H_L score is the more perceptually relevant the given outlier is. The plot shows outliers detected for the unit selection costs.

again the condition $S_2(i) \geq thr$. The *Missed Rate* has been defined as follows:

$$Missed\ Rate = \frac{N_{mis}}{N_{annot}} \times 100\,[\%]\,, \tag{5.8}$$

where N_{mis} is a number of *annotated artifacts* in a close vicinity[6] of which no outlier has been found, and N_{annot} is a total number of *annotated artifacts.*

Interestingly enough, when the outliers of all components are combined, the actual *Missed Rates* are 36.71 % and 72.84 % for the `Syst1` and the `Syst2`, respectively. This means that the outliers found in the costs of the optimal sequences of the units forming the outputs of the `Syst1` significantly better reflect the locations of the audible artifacts than the outliers found for the

[6] Note that the close vicinity was defined for this experiment as 2 phonemes from a likely artifact location estimated manually by the author based on the listeners' annotations. To automate the evaluation procedures, this setting will be changed to an exact phoneme match for further experiments described in Sec. 5.4.

Tab. 5.1: Perceptual relevance of outliers—concatenation cost sub-components.

	En		F0		MFCC	
	`Syst1`	`Syst2`	`Syst1`	`Syst2`	`Syst1`	`Syst2`
Hit Rate [%]	31.25	57.89	80.00	90.91	45.45	60.00
Missed Rate [%]	82.98	86.42	91.49	87.65	89.36	96.30

Tab. 5.2: Perceptual relevance of outliers—unit selection costs.

	ConC		TgtC		TotC	
	`Syst1`	`Syst2`	`Syst1`	`Syst2`	`Syst1`	`Syst2`
Hit Rate [%]	83.33	77.78	40.79	50.00	33.33	100.00
Missed Rate [%]	89.36	91.36	59.57	98.77	91.49	96.30

`Syst2`.

Note that a detailed analysis of the differences between the two systems is out of the scope of this thesis as further experiments will only be done with the `Syst2` due to its feasibility for modifying its search algorithm.

5.3.3 Trimming off the Outliers

Having obtained the results of the annotation experiment, it was interesting to speculate how the quality of the synthetic utterances change (if at all) when a limit is set for the costs during searching for the optimal sequences of units in the unit selection. The role of the limit would be to avoid the selection of units of the outlying costs by pruning the beam.

Obviously, too radical pruning of the search space can lead to the inability of the search algorithm to find the target sequences of units. Nevertheless, having a large unit database on hand, such an experiment can be conducted. Each concatenation cost sub-component, as well as the costs themselves, were assigned a maximum threshold which was equal to the value of the upper whiskers of the respective boxplots shown in Fig. 5.1. The test sentences were

Tab. 5.3: Impact of introducing a limit for the unit selection costs and the concatenation cost sub-components

	En	F0	MFCC	Join Cost	TgtCost	TotCost
Improvement [%]	31.25	41.67	33.33	50.00	66.67	66.67
Deterioration [%]	18.75	16.67	16.67	10.00	0.00	33.33
No impact [%]	50.00	41.67	50.00	40.00	33.33	0.00

re-synthesized using the thresholds. In none of the cases, the trimming off the outliers lead to the synthesis failure.

To evaluate the impact of the modification of the search algorithm, the ABX preference test was conducted. The original and the re-synthesized sentences were presented to listeners in randomized pairs. The task of the listeners was to express their preference regarding the overall quality of the presented samples. The test also contained sentences that were identical due to not containing any outliers. These sentences were used to check the reliability of the ratings as no preference was expected for the pairs containing them.

The following results were obtained from 9 listeners in total: 5 preferences for the original sentences, 9 sentences with no preference and 10 preferences for the sentences generated by the modified system. Note that the figures represent ratings for which 60 % of listeners found an agreement, also pairs containing identical sentences are not included. Strictly speaking, the obtained result does not allow to statistically reject a null hypothesis that the original system is equally good or better than the modified system. Nevertheless, a slight preference to the modified system exists, which makes it interesting to analyze the removal of which outliers lead to the largest improvement rate. The result of this analysis is shown in table Tab. 5.3.

5.3.4 Discussion

Some agreement can be found for the results of the two perceptual experiments—the annotation and the preference listening tests. At the same time, differences also exist.

Let us first take a look at the results obtained for the target cost. It can be seen that removing the related outliers leads to improvements of the quality of the output of the system in 66.67 %. This is in contrast to the perceptual importance of the target cost outliers obtained in the first test, which was found to be 50.00 % in terms of the *Hit Rate*. A hypothesis is that this is due to the different nature of the two perceptual experiments. While the first one implicitly pose a requirement on the listeners to mark as short segments as possible, the target cost would actually require the opposite as it is rather a supra-segmental cost. On the other hand, setting a limit for the target cost has bigger effect on the behavior of the system. This is because the target cost outliers appear in larger quantities due to largely binary nature of the cost, and also because when the target cost is “violated”, the unit selection stays with this “violation” as long as the concatenations are believed to be smooth according to the concatenation cost or not needed at all. The hypothesis will further be discussed in Sec. 5.4.2, where results of additional experiments with the outliers detection approach are presented.

If we turn next to the concatenation cost and its sub-components, it can be seen that better consistency was found between the two experiments. With respect to the discussion in the previous paragraph, this is perfectly understandable result. Also, in light of the results presented in Chapter 3, as no surprise comes the finding that $F0$ is the most important component of the concatenation cost in its current implementation, even though it only measures the static $F0$ differences at concatenation points.

5.3.5 Conclusions

To conclude, it has been shown that a small system improvement can be achieved by setting a limit for the costs during the search of the optimal sequences of units if the size of the unit database allows for pruning the search space.

Some additional questions however arise. First, to what extend the limit for the costs can be lowered, and what the impact on the quality of the system output would be. Second, as the large number of annotated artifacts do not co-occur with the detected outliers (72.84 % for `Syst2`), would lowering of the threshold for the outliers detection improve the match, and if not, how many of the annotated artifacts can be attributed to concatenations. In the following sections, these questions will be addressed.

5.4 Tuning the Analytic Method

5.4.1 Tunable Parameters

Some parameters used in the pilot experiment with the new analytic method based on the cost outlier detection described in the previous section were set ad hoc. In particular, three parameters were set—the summation length for the listeners ratings (set as $L = 2$ for the H_L *score* (5.3)), the span of the whiskers—*outlier detection threshold*—for the identification of the cost outliers (set as $1.5 \times IQR$[7], Fig. 5.1), and a *perceptual relevance threshold*, i.e. the threshold for deciding whether or not an outlier is perceptually relevant (set as $thr = 0.5$ for the sum S_2 (5.6) of the H_L *scores*).

The setting of the length L of the summation window can be justified by the observation that the listeners are inclined to use 3–5 phoneme long segments for annotating the audible unit selection artifacts. The other two parameters can however be tuned with the objective to maximize the *Hit Rate* (5.7) and minimize the *Missed Rate* (5.8).

[7] IQR=Inter-Quartile Range

Outlier Detection Threshold

It can be seen in Fig. 5.5-5.6 that more audible artifacts can potentially be hit by lowering the threshold for detecting the outliers. At the same time, the lower the threshold is the more likely it is that the search algorithm will not be able to deliver any sequence of target units in the pruned space. It is also clear that lowering the threshold leads to a higher number of false alarms.

In order to get more insight into the sensitivity of the outlier detection, four different thresholds were experimented with—$0 \times IQR$, $0.5 \times IQR$, $1.0 \times IQR$ and $1.5 \times IQR$.

Perceptual Relevance Threshold

Many phonemes can theoretically be identified as locations of audible artifacts in the annotation experiment described in Sec. 5.3.2. Obviously, the more annotators are invited to provide annotations, the more likely any phoneme in the synthesized sentences becomes a location of an audible artifact due to being annotated by at least one of the listeners. This is explained by the subjectivity of the TTS quality evaluation as well as the uncertainty about the exact locations of the artifacts.

The H_L *score* (5.3) was introduced to measure the perceptual relevance of the detected outliers. This score can actually be used to assign a value to every phoneme in the test data, no matter if it coincides with an outlier location or not. As already explained, it is beneficial for better robustness to use the summation $S_2(i)$ of the H_L *scores.*

In order to define different values of the thresholds for the perceptual relevance that should be used in the tuning, the $S_2(i)$ value was assigned to all phonemes in the annotated test data. The phonemes were then ranked according to the assigned values, and the thresholds have been defined as 20th, 40th, 60th, 80th and 100th percentiles. By increasing the threshold, only phonemes that were identified as the locations of artifacts by a large

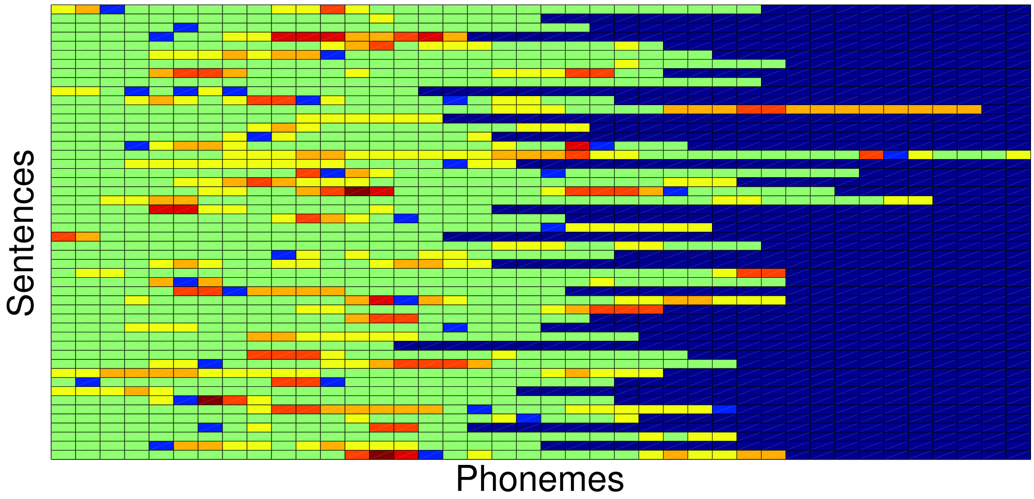


Fig. 5.5: Outlier positions in annotated data - whisker position $1.5 \times IQR$.

The figure shows annotations of artifacts similarly to Fig. 5.2 complemented with positions of outliers (light blue squares) detected using a default outlier detection threshold. Note that the test sentences have different lengths, navy blue is used as a figure background.

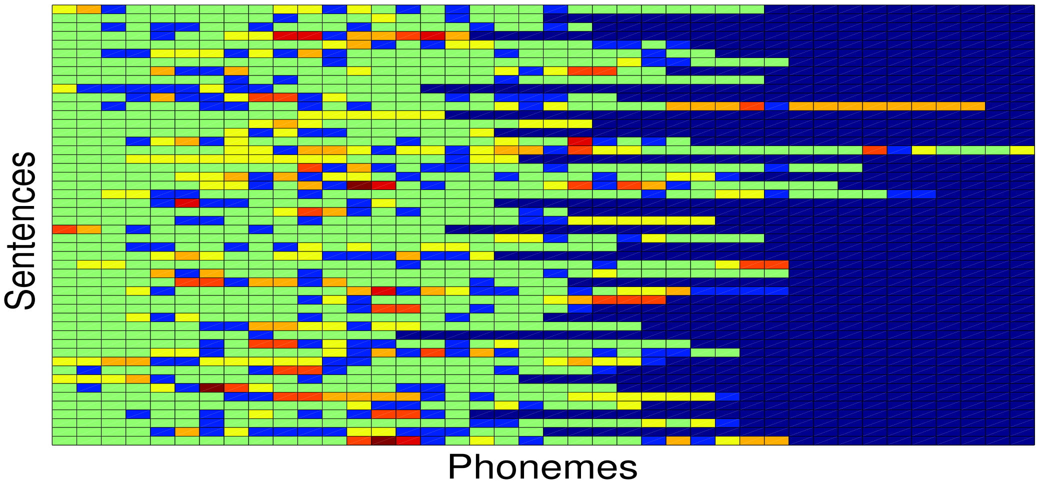


Fig. 5.6: Outlier positions in annotated data - whisker position $0 \times IQR$.

The figure shows annotations of artifacts similarly to Fig. 5.2 complemented with positions of outliers (light blue squares) detected using a low outlier detection threshold. Note that the test sentences have different lengths, navy blue is used as a figure background.

number of listeners are taken into account. For instance, the 20th percentile would only take into account 20 % of the most frequently annotated artifacts.

5.4.2 Results

Cost Comparisons

We turn first to the comparison of the individual costs. In Fig. 5.7-5.10, the contours of the *Hit Rate* vs. *Missed Rate* measure for all costs and the non-additive combination of the costs and the concatenation cost sub-components (denoted as `All` in the plots) are shown. The curves are obtained by changing the *perceptual relevance threshold*. Each plot corresponds to one setting of the *outlier detection threshold*.

It is obvious that the lower percentile is used for the *perceptual relevance threshold*, the lower the *Hit Rate* is. This is because the number of the detected outliers under a given *outlier detection threshold* is constant whereas the total number of *annotated artifacts* is decreasing. Ideally, a cost that scales well with the listeners' annotations would be represented by a "flattening" curve, i.e. the decrease in the *Hit Rate* would be slower than the decrease in the *Missed Rate*. In an extreme case, when the obtained curve is a horizontal line, all detected outliers for a given cost would be highly perceptually relevant. Vertical line would represent a cost, the outliers of which do not discriminate the perceptual importance of the artifacts.

It can also be seen in the figures, how the individual costs combine together. For instance, in Fig. 5.7, the target cost outliers contribute to the concatenation cost outliers which then demonstrates as a shift of the total cost curve to the right. For comparison, a non-additive cost composed of all individual costs and the concatenation cost sub-components is shown to demonstrate how all the costs and the sub-components combine.

Several observations can be made from the figures. First, as already observed in the pilot experiment described in Sec. 5.3, the target cost is the least

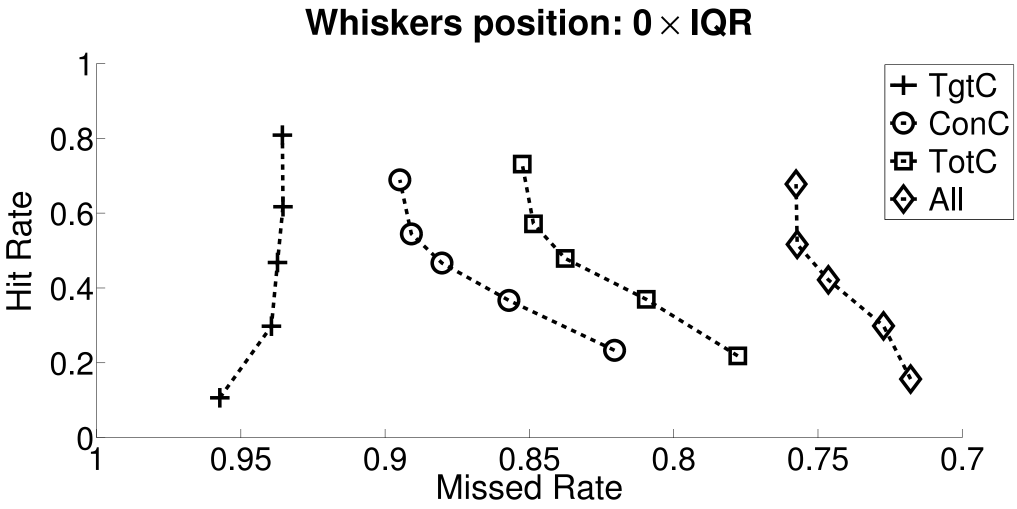


Fig. 5.7: Comparison of the perceptual relevance of cost outliers—0 × IQR

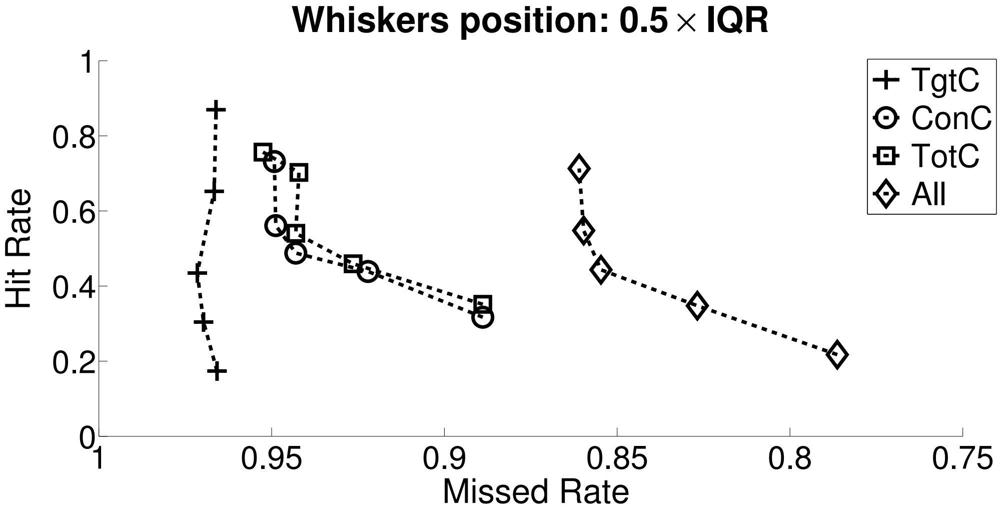


Fig. 5.8: Comparison of the perceptual relevance of cost outliers—0.5 × IQR

The points of the individual contours represent values calculated for different perceptual relevance thresholds—100th, 80th, 60th, 40th and 20th percentiles as defined in Sec. 5.4.1. `All` *stands for a non-additive combination of all costs and the concatenation cost subcomponents.*

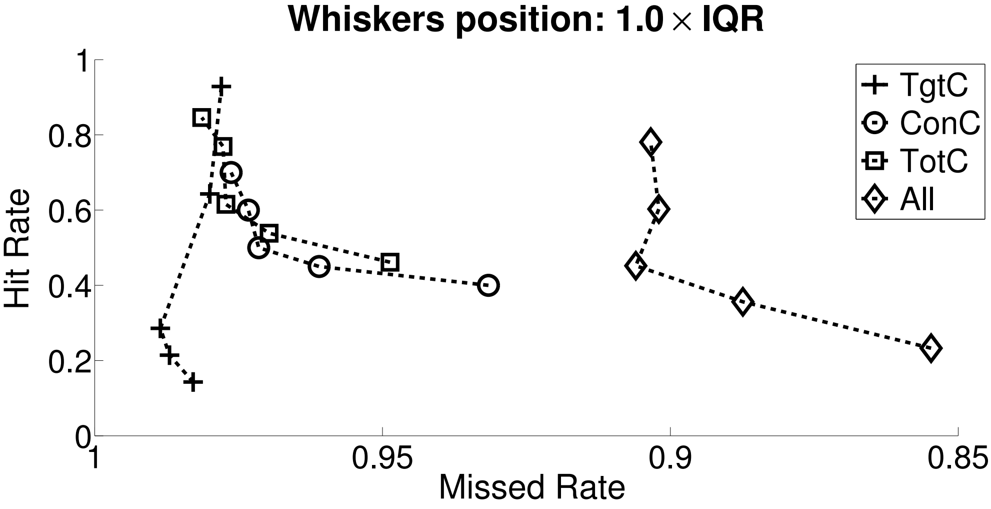


Fig. 5.9: Comparison of the perceptual relevance of cost outliers—1.0 × IQR

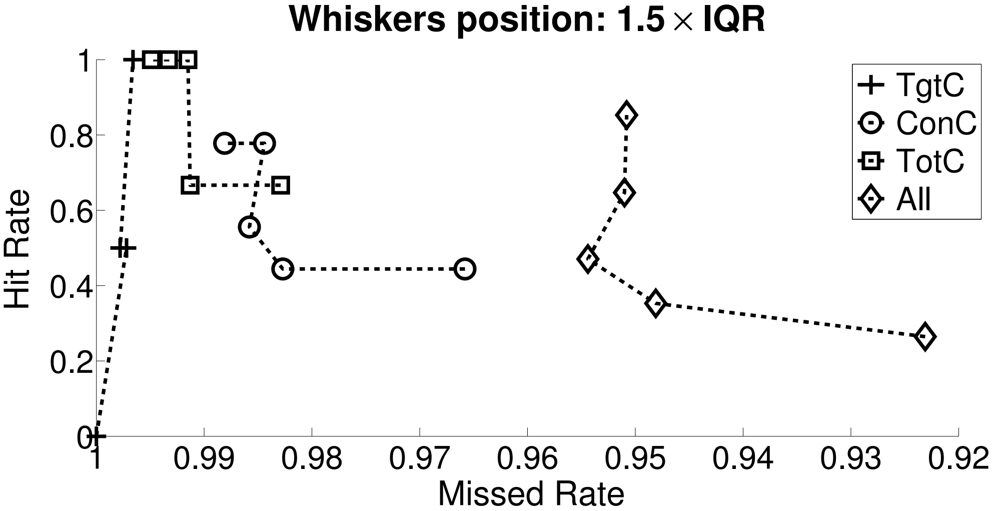


Fig. 5.10: Comparison of the perceptual relevance of cost outliers—1.5 × IQR

The points of the individual contours represent values calculated for different perceptual relevance thresholds—100th, 80th, 60th, 40th and 20th percentiles as defined in Sec. 5.4.1. `All` *stands for a non-additive combination of all costs and the concatenation cost subcomponents.*

corresponding one to the audible artifacts. In Sec. 5.3.4, it was hypothesized that this comes as a result of the nature of the design of the annotation experiment. Nevertheless, if we look at Fig. 5.7, it can be seen that, in contrast to Fig. 5.8-5.10, the target cost outliers complement the concatenation cost outliers as the total cost curve is shifted to the right. This would suggest that the current target cost implementation in the `Syst2` is lacking sufficient perceptual discriminativeness.

Second observation is related to the comparison of the concatenation and the total costs. If we look at the curves obtained for the *outlier detection thresholds* $0.5 \times IQR$ and $1.0 \times IQR$, it can be seen that the concatenation cost shows comparable or even slightly better performance. This would mean that the total cost outliers are largely given by the concatenation cost outliers, and the target cost component of the total cost is rather, albeit negligibly, “masking” the audible artifacts. Fig. 5.10 depicting the comparison for the *outlier detection threshold* $1.5 \times IQR$ however shows that this “masking” effect may be of benefit when the pruning of the search space as suggested in Sec. 5.3.3 can only be conservative, and avoiding false alarms is of interest. Under this setting, pruning the search space only based on the total cost values would be more conservative than based on the concatenation cost values because the curve for the total cost represents higher *Hit Rates*.

Last but not least, if a comparison of any of the costs with the non-additive combination of all costs and the concatenation cost sub-components is made, a clear shift to lower *Missed Rates* can be seen. Unfortunately, for the *outlier detection threshold* $0 \times IQR$, the non-additive combination loses its “flatness” due to the target cost component.

Comparison of Concatenation Cost Sub-Components

In this paragraph, plots showing a comparison of the perceptual relevance of the outliers detected for the individual concatenation cost sub-components are presented.

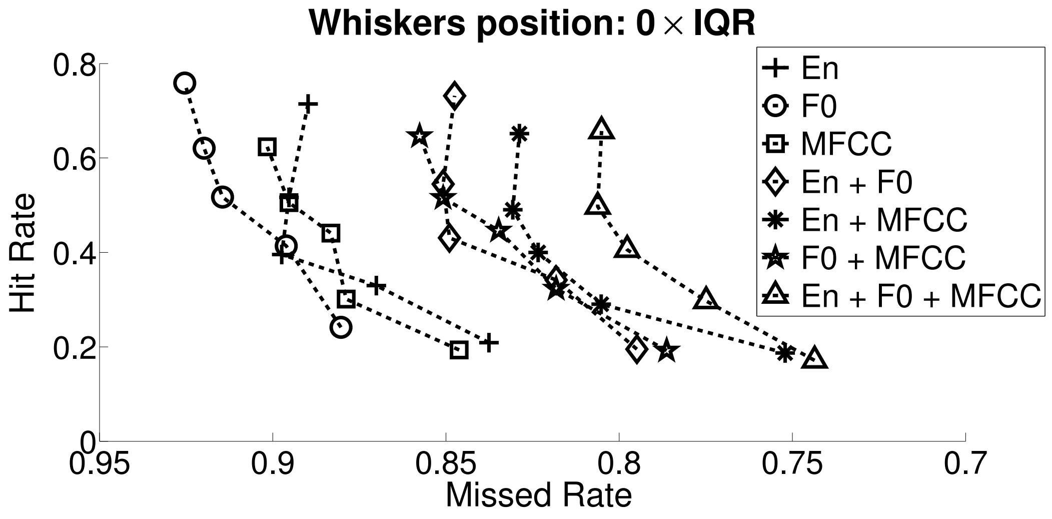


Fig. 5.11: Comparison of the perceptual relevance of concatenation cost sub-component outliers—0 × IQR

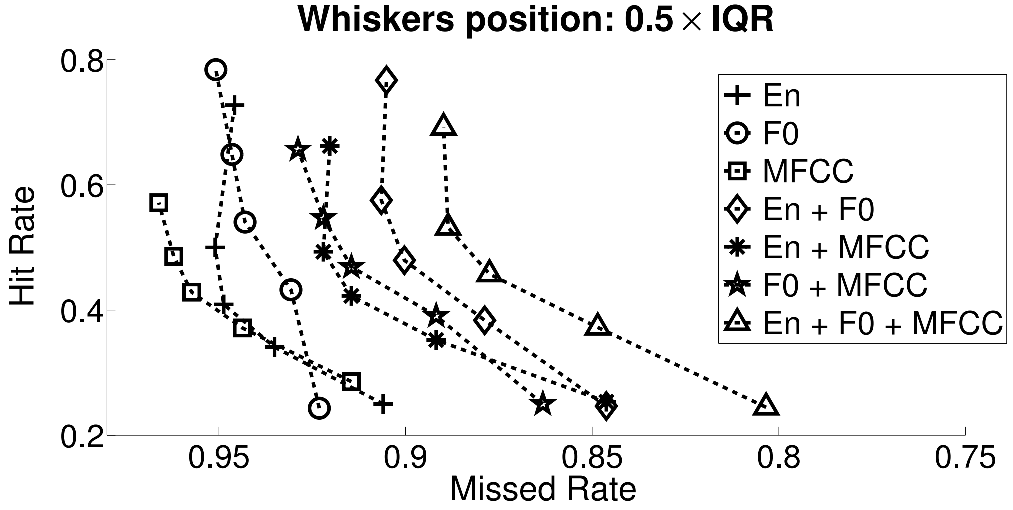


Fig. 5.12: Comparison of the perceptual relevance of concatenation cost sub-component outliers—0.5 × IQR

The points of the individual contours represent values calculated for different perceptual relevance thresholds—100th, 80th, 60th, 40th and 20th percentiles as defined in Sec. 5.4.1.

Fig. 5.11-5.12 show the curves obtained for the *outlier detection thresholds* $0 \times IQR$ and $0.5 \times IQR$. It can be seen that all sub-components show similar behavior and by their non-additive combinations, the curves shift to the right. At the same time, the *Hit Rates* go down to values around 0.2, which means that the detected outliers are in many cases false alarms.

By increasing the *outlier detection threshold*, i.e. the $1.0 \times IQR$ and $1.5 \times IQR$ settings shown in Fig. 5.13-5.14, the sub-components start to behave differently. For the $1.0 \times IQR$, the $F0$ sub-component does not show a good perceptual discriminativeness, in contrast to the energy sub-component. Its performance is however improved by increasing the *outlier detection threshold* to $1.5 \times IQR$. This is the motivation for analyzing the behavior of the costs and the concatenation cost sub-components with respect to different *outlier detection thresholds* as a sensitivity analysis described in the next paragraph.

Second observation worth mentioning is the large *Missed Rate* of the MFCC sub-component. As a result, the combination of the energy and the $F0$ sub-components show a curve similar to the combination of all the sub-components.

Sensitivity Analysis

In order to show the sensitivity of outliers of each cost and the concatenation cost sub-components with respect to the audible artifacts, Fig. 5.15-5.20 give curves obtained across the *outlier detection thresholds*. The plots show what thresholds should be used to leverage the information obtained in the outliers in the best way.

Obviously, the most aggressive setting of the *outlier detection threshold* could be used for the unit selection cost limits. This would, on one hand, allow for potential removing of a larger number of audible artifacts, but on the other hand, it can also lead to an inability of the unit selection search algorithm to deliver a target sequence of units due to too aggressive pruning of the search space. The other extreme is to use the most conservative setting, i.e.

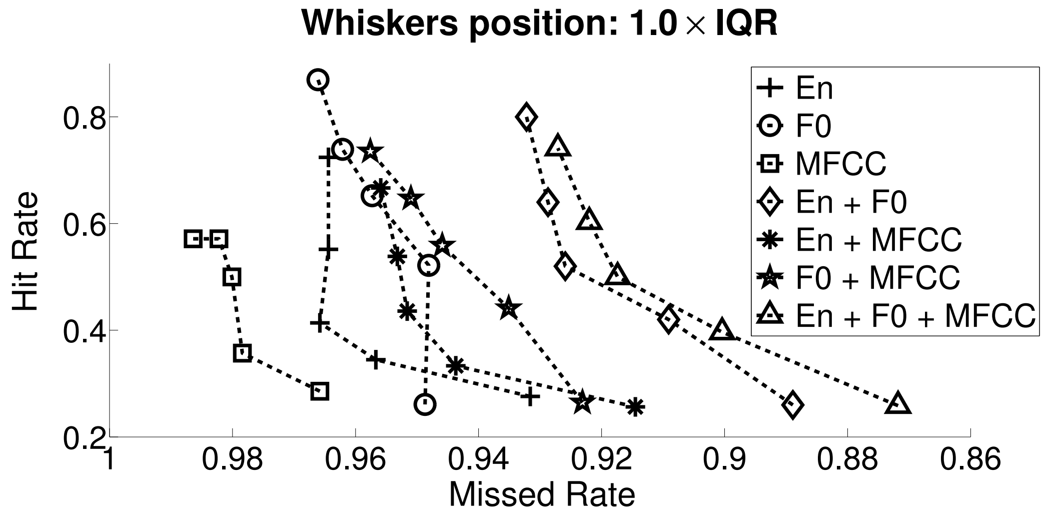


Fig. 5.13: Comparison of the perceptual relevance of concatenation cost subcomponent outliers—1.0 × IQR

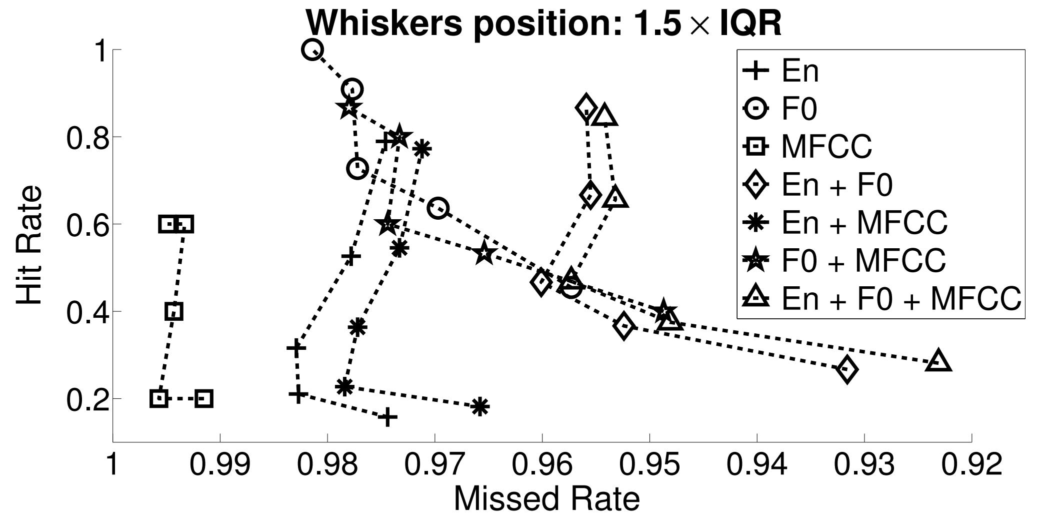


Fig. 5.14: Comparison of the perceptual relevance of concatenation cost subcomponent outliers—1.5 × IQR

The points of the individual contours represent values calculated for different perceptual relevance thresholds—100th, 80th, 60th, 40th and 20th percentiles as defined in Sec. 5.4.1.

$1.5 \times IQR$. In this case a low number of outliers is found, and as a result there is less potential for removing the audible artifacts. The results of a perceptual experiment using these two settings are presented in Sec. 5.4.3. Based on the sensitivity analysis presented in this paragraph, the best setting would however differ for the individual costs and the concatenation cost sub-components.

For the concatenation cost, the good setting would be $1.0 \times IQR$. Setting a limit for the target cost is not of much benefits as it leads to a large number of false alarms for any setting. Finally, one can choose between the $1.0 \times IQR$ and $0.5 \times IQR$ settings for the total cost limit.

If we turn next to the concatenation cost sub-components, all settings for energy lead to a considerably large number of false alarms. If the database size allows, the recommended setting would however be the most aggressive one. For the $F0$ sub-component in its current implementation, the least aggressive threshold appears to be the best choice. For the MFCC sub-component, the setting $0.5 \times IQR$ leads to reasonable results, although similarly to the energy sub-component, false alarms are frequent under all settings. The impacts, both perceptual and algorithmic, of introducing the limits based on these settings will also be presented in Sec. 5.4.3.

For the sake of completeness, Fig. 5.21 presents the sensitivity analysis for the combination of all costs and the concatenation cost sub-components. It can be seen that under all thresholds, the frequency of false alarms is similar. This is the consequence of the false alarms found for the individual components.

It is also notable that even under the most aggressive setting, more than 70 % of the audible artifacts remain undetected by the outliers. We will further comment on this finding below in Sec. 5.4.4.

5.4.3 Improving TTS Quality

In Sec. 5.3.3, the impact of setting a conservative limit for the unit selection costs and the concatenation cost sub-components was evaluated using the ABX preference test. The same test was used to evaluate the impact of two other

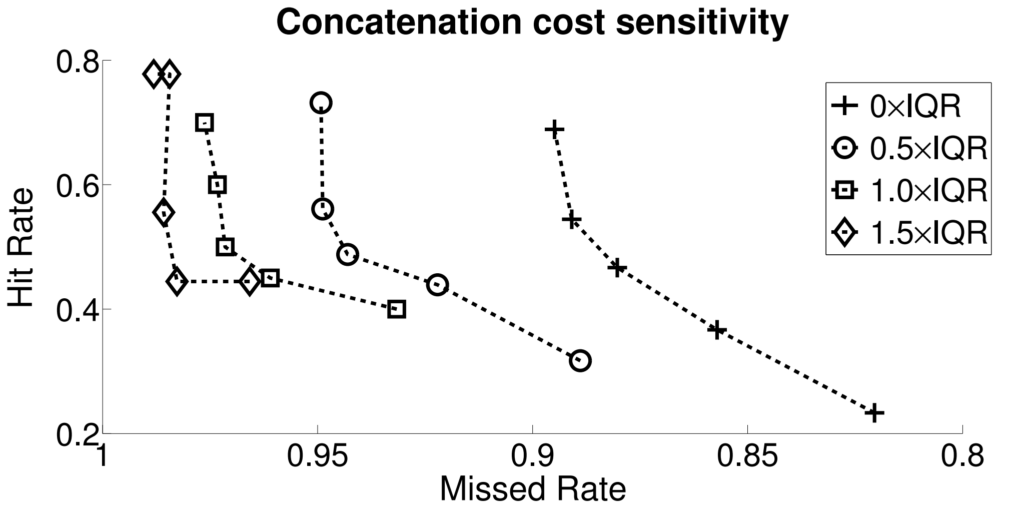


Fig. 5.15: Outlier sensitivity analysis—concatenation cost

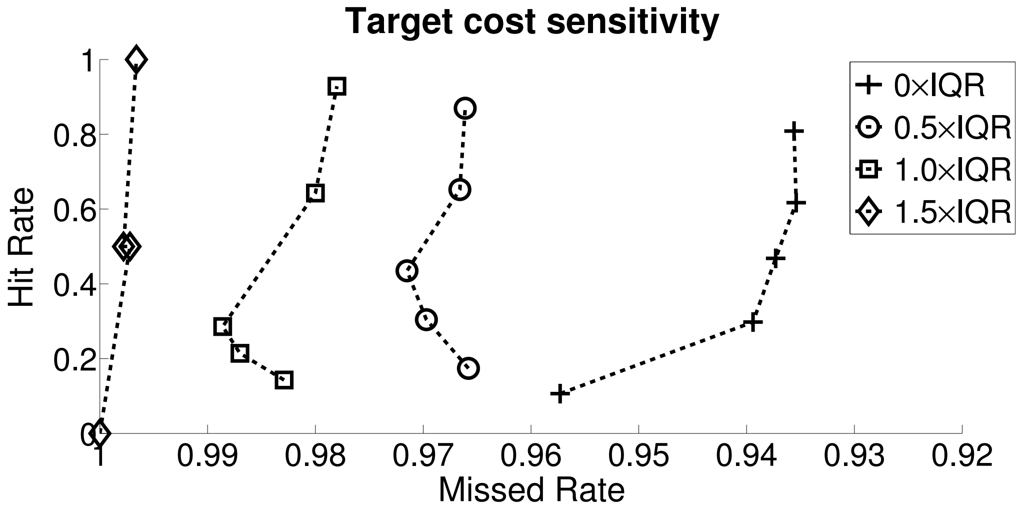


Fig. 5.16: Outlier sensitivity analysis—target cost

The points of the individual contours represent values calculated for different perceptual relevance thresholds—100th, 80th, 60th, 40th and 20th percentiles as defined in Sec. 5.4.1.

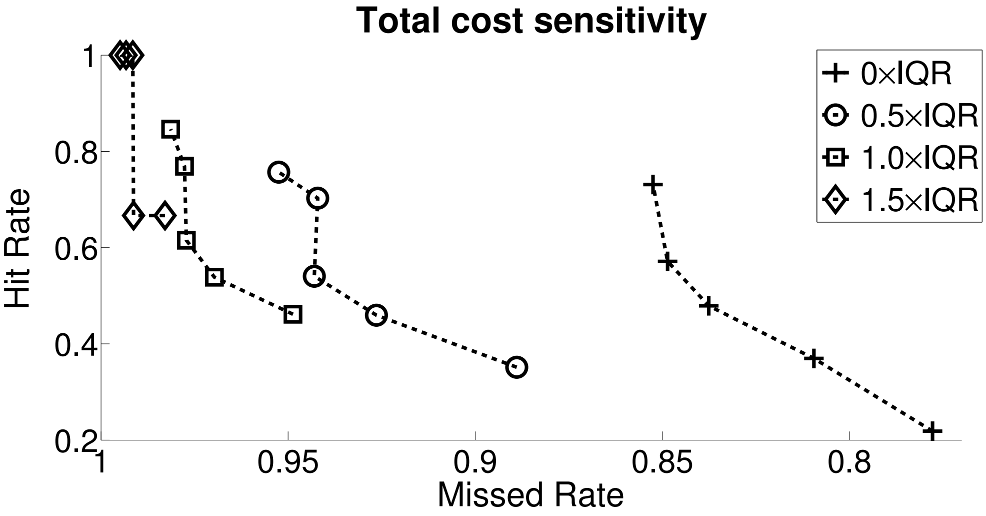


Fig. 5.17: Outlier sensitivity analysis—total cost

The points of the individual contours represent values calculated for different perceptual relevance thresholds—100th, 80th, 60th, 40th and 20th percentiles as defined in Sec. 5.4.1.

settings—the most aggressive one, i.e. $0 \times IQR$ (henceforth referred to as `Aggr`) and a setting differentiating the individual costs and the concatenation cost sub-components (henceforth referred to as `Diff`). The `Diff` setting was defined in line with the description given in Sec. 5.4.2. For the total cost, $0.5 \times IQR$ was used as the *outlier detection threshold*.

The results of the evaluation are summarized in Tab. 5.4. For the sake of completeness, also the results obtained for the conservative setting $1.5 \times IQR$ (referred to as `Cons`) already presented in Sec. 5.3.3 are repeated in the table.

Unfortunately, for both the `Aggr` and the `Diff` settings, the synthesis failures occurred as the search space was pruned too much. For the `Aggr`, it happened in 31 (62 %) cases, for the `Diff` setting, in 4 (8 %) cases. An interesting observation is that only for the `Cons` setting some quality gain has been achieved. This finding will be discussed in the next subsection.

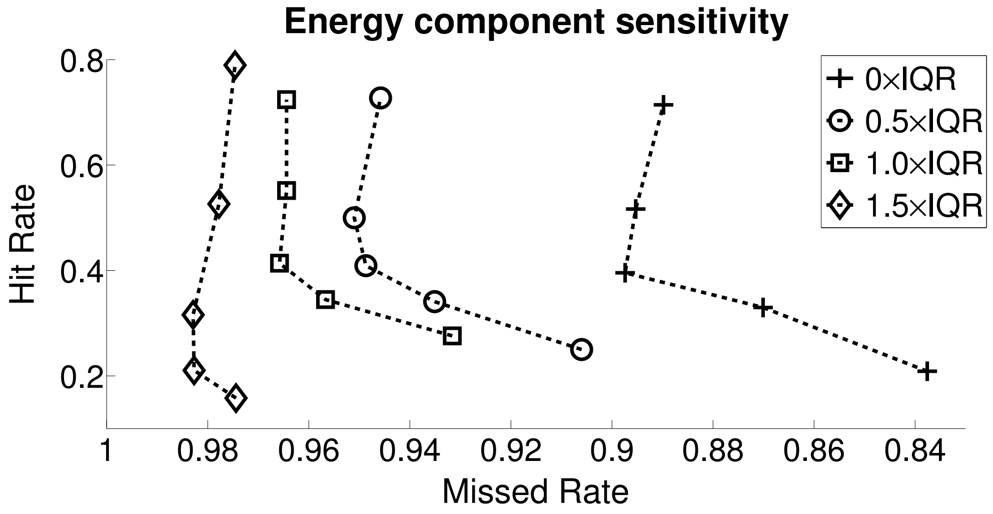


Fig. 5.18: Outlier sensitivity analysis—energy sub-component

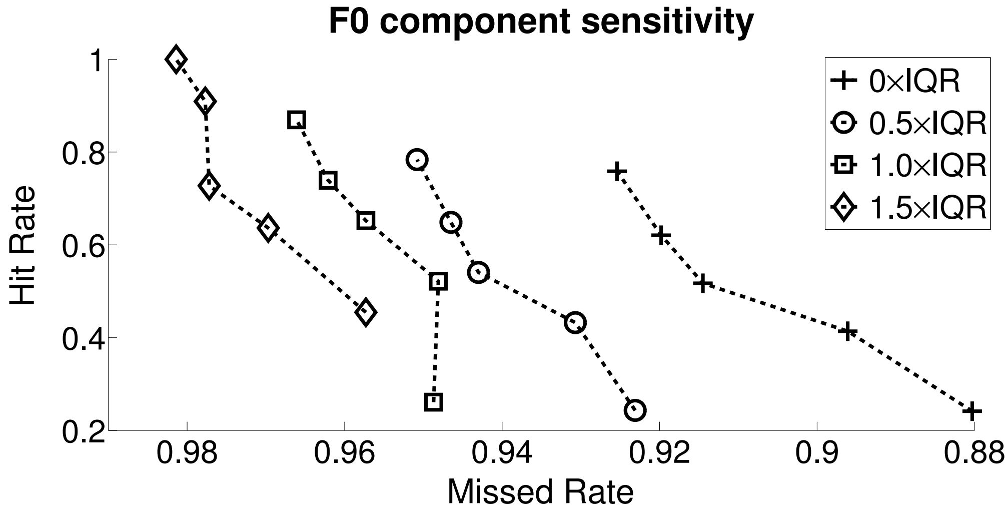


Fig. 5.19: Outlier sensitivity analysis—F0 sub-component

The points of the individual contours represent values calculated for different perceptual relevance thresholds—100th, 80th, 60th, 40th and 20th percentiles as defined in Sec. 5.4.1.

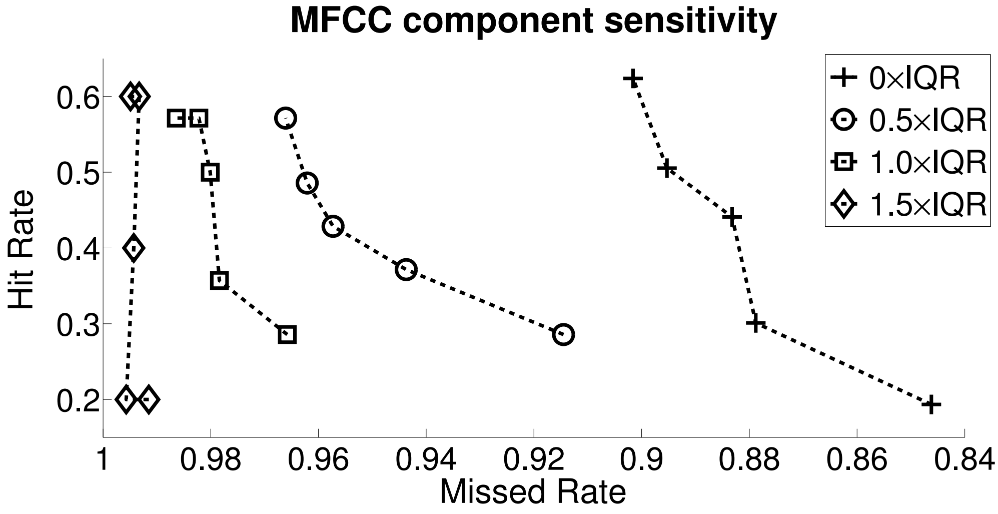


Fig. 5.20: Outlier sensitivity analysis—MFCC sub-component

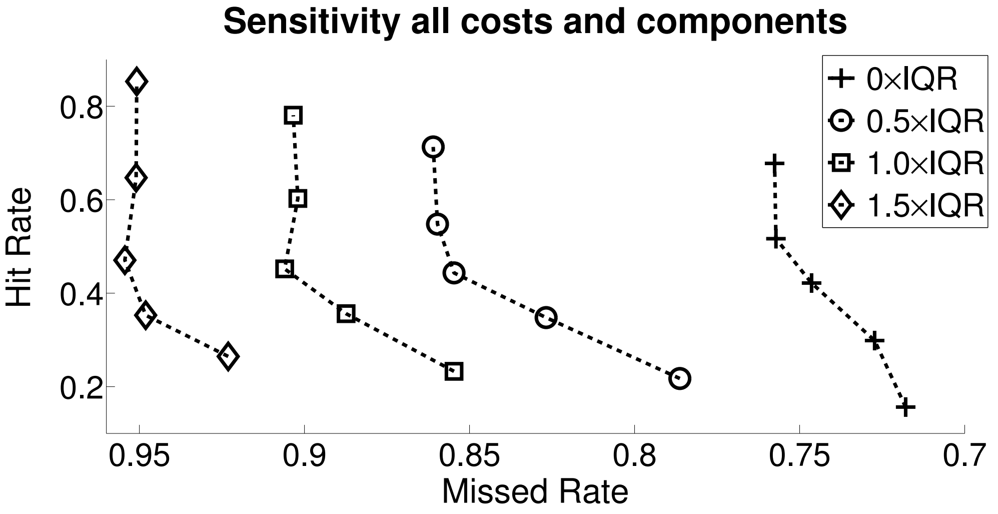


Fig. 5.21: Outlier sensitivity analysis—a combination of all costs and concatenation cost sub-components

The points of the individual contours represent values calculated for different perceptual relevance thresholds—100th, 80th, 60th, 40th and 20th percentiles as defined in Sec. 5.4.1.

Tab. 5.4: Impact of setting limits for the unit selection costs and the concatenation cost sub-components

Setting	PrefOrig	PrefMod	No Pref	Num Sent	No Diff
`Cons`	5	10	9	50	21
`Aggr`	4	5	6	19	2
`Diff`	9	6	13	46	2

5.4.4 Discussion

The finding that the system output quality cannot be much improved by setting strict limits for the unit selection costs and the concatenation cost sub-components has not been entirely surprising to us as the *Missed Rates* of all outlier groups and also their combinations are high. Still, the obtained results have been disappointing.

Let us speculate at this point about the possible reasons for this. First and the most obvious reason could be the low perceptual relevance of the system costs and also the concatenation cost sub-components. As a consequence, the system may behave slightly randomly. Nevertheless, at least for the most extreme values, the costs reflect human perception. This can be concluded from the preference test evaluating the `Cons` setting.

Despite the big size of the system unit inventory, the second possible reason could be data scarcity. It is possible that for some specific contexts, good candidate units cannot simply be found. This hypothesis would be supported by two observations. First, the `Aggr` setting leads to synthesis failures, even though it keeps in average approximately 75 % of candidates. Second, even in cases when a sentence can be synthesized under this aggressive setting, the audible artifacts remain in the system outputs.

5.4.5 Conclusions

The experiments aiming at tuning the analytic method based on the outlier detection have shown that the *Missed Rates* can be lowered by the tuning.

At the same time many false alarms are introduced by the more aggressive settings. This suggests that the perceptual relevance of the currently used costs and the concatenation cost sub-components is limited. More precisely, the correspondence of their extreme values and the audible artifacts haphazardly appearing in the unit selection outputs is low.

The distribution of the values of the target cost, the MFCC and the energy concatenation cost sub-components have a large dispersion. This makes it difficult to identify deviating values corresponding to the audible artifacts. For the $F0$ sub-component, this separation is possible. There is however only few deviating values.

These observations clearly show that in order to be able to reliably identify and possibly also to avoid audible unit selection artifacts, there is still a lot of work to be done in the design of the costs and their sub-components. The proposed analytic method can however help in this research. The guidance on using the method will be provided in Sec. 5.6.

5.5 Role of Concatenation Artifacts

5.5.1 Motivation

It has been shown that a large number of the audible artifacts cannot be detected by the outliers of any of the costs or the concatenation cost sub-components. In order to understand what quality improvement can potentially be achieved by a new concatenation cost function, it needs to be analyzed how many of the undetected audible artifacts appear at concatenation points.

5.5.2 Procedure

In Sec. 5.3.2, two quality metrics have been introduced to measure how well the outliers predict audible artifacts—*Hit Rate* (5.7) and *Missed Rate* (5.8). In fact, the outliers represent pointers to certain phonemes in synthesized sentences. The same can be said about concatenation points as the concatenations are done in the middle of phonemes. Therefore, if the outlier positions are replaced by positions of the concatenation points, and the same calculations as described in Sec. 5.4.1 are done, a plot similar to Fig. 5.7-5.21 can be obtained. From this plot, an estimate can be made of how many of the audible artifacts are due to bad concatenations.

5.5.3 Results

Fig. 5.22 shows the result of the experiment. The gap between the curve obtained for the outliers of the concatenation cost function (`ConC`) and the curve of its sub-components (`En + F0 + MFCC`) shows what can be gained by a non-additive implementation of the concatenation cost function. The gap between the curves obtained for the outliers of the concatenation cost sub-components (`En + F0 + MFCC`) and for the concatenation points (`Concat`) gives an estimate of how much potential still exists for improving the concatenation cost function.

Based on the obtained result, an ideal concatenation cost function should achieve a *Missed Rate* of approximately 50 % for the most perceptually relevant artifacts (the lowest point of the `Concat` curve), but unlike the `Concat` curve, its curve should be flat, i.e. its *Hit Rate* should stay high.

The `Concat` curve also shows another interesting outcome of the experiment. Based on the *artifact annotations*, approximately 40 % of the concatenation points are never identified as locations of the audible artifacts. This is explained by the *Hit Rate* of approximately 60 % when taking into account all listeners' annotations (the highest point of the `Concat` curve). At the same time, only 20 % of the concatenation points correspond to the most perceptually relevant artifacts. This is represented by the lowest point of the `Concat` curve corresponding to the 20th percentile. These figures can be used for estimating the probability of a concatenation artifact per sentence, which can serve as a quality metric for comparing different unit selection systems.

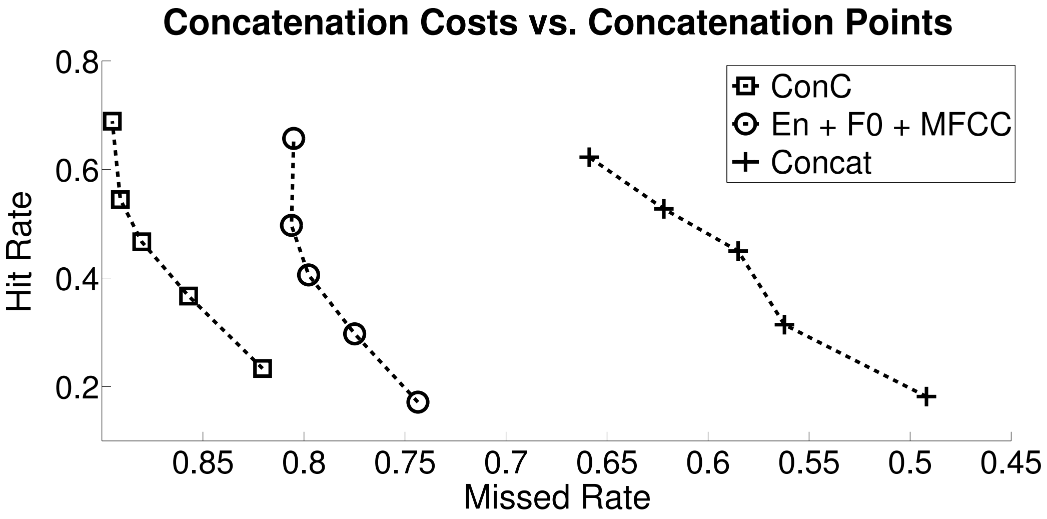


Fig. 5.22: Perceptual relevance—comparison of concatenation costs and concatenation points

> *The figure shows comparison of the perceptual relevance of outliers of the concatenation cost, its subcomponents (both calculated at* $0 \times IQR$ *outlier detection threshold) and concatenation points (*`Concat`*). The points of the individual contours represent values calculated for different perceptual relevance thresholds—100th, 80th, 60th, 40th and 20th percentiles as defined in Sec. 5.4.1.*

5.5.4 Discussion

Fig. 5.21 has shown that more than 70 % of the audible artifacts are not detected by any of the outlier groups even under the most aggressive outlier detection setting. Fig. 5.22 shows that the potential improvement that can be achieved by introducing a new concatenation cost function is approximately 25 %.

This means that there are still a lot of audible artifacts that are not explained. These audible artifacts can either be identified by an improved target cost function or they can also represent artifacts that are already present in the unit inventory (e.g. signal distortions, smacks, etc.). Their nature can be estimated by investigating their locations or, more precisely, by investigating whether they appear in sequences, which would suggest that they can likely be detected by a target cost function, or in isolation, which would rather point to problems related to recordings.

Also, the real potential of the concatenation cost function improvement is probably lower than the estimate as it can be expected that some of the bad concatenations won't be possible to detect due to mislabeling of units (e.g. pitch tracking errors). A detailed analysis of these sources again remains for our future work.

5.6 Guidance on the Analytic Method

The analytic method introduced in the previous sections can generally be used for analyzing any unit selection based TTS system. Moreover, it can also be used to some extent for tuning the unit selection costs and/or testing new cost functions without a need to conduct costly listening tests.

The procedure for evaluating a new cost function can be summarized as follows:

1. Synthesize a random set of sentences by a unit selection based TTS system.
2. Annotate audible artifacts present in the synthesized sentences so that they can be related to the system base units.
3. Identify outliers of an original cost function of interest and calculate *Hit Rates* (5.7) and *Missed Rates* (5.8).
4. Introduce a new cost function and resynthesize the set of sentences while keeping the original sequences of units fixed.
5. Calculate *Hit Rate* and *Missed Rate* of the new cost function and compare them to the original ones. From the difference in the rates, performance of the new cost function can be seen.

5.7 Conclusions and Future Work

In this chapter, a detailed analysis of the current implementation of the unit selection method in the ARTIC TTS system has been presented. The analysis has been focused on the system costs and the concatenation cost subcomponents.

It has been shown that the current costs and the concatenation cost subcomponents do not correspond well to audible artifacts present in the system outputs as annotated by human judges. Approximately 70 % of the audible artifacts are missed by the extreme values of the costs or the concatenation cost sub-components. As a consequence, the system output quality can only slightly be improved by avoiding the selection of units of very extreme cost values. It is also possible that only moderate system improvement would be achieved by having better costs as the data scarcity problem may play a role as well.

It has also been shown how many of the audible artifacts coincide with locations of concatenation points, and an estimate has been made of how big quality improvement can potentially be achieved by introducing a new concatenation cost function or lowering the number of concatenations. The quality gain is estimated as 25 % percent in terms of a number of the audible artifacts.

The analysis was done by applying a new method that has also been proposed in this chapter. The method is not constrained to the TTS system considered in this thesis. It can be used for analyzing any unit selection based TTS system. A guidance has been provided on how to leverage the method for testing new cost implementations without a need to conduct listening tests.

Chapter 6

Conclusions

The main objective of this thesis was to design a new concatenation cost function for the ARTIC TTS system. Already from the literature review presented in Sec. 1.5, it can be seen that measuring concatenation artifacts in the unit selection based synthesis is a traditional problem. Many researchers have tried to solve this problem, but with little success.

Our original aim was to investigate concatenation points across different phonemes. We have started our analysis by looking at short Czech vowels as the concatenations in vowels are the most salient ones. It has turned out that the problem of measuring the concatenation discontinuities in mid-vowel joins is very complex. This explains why so much efforts have already been spent in this area over the last one and a half decades, and still we do not have a clear understanding of all latent phenomena influencing human perception of the discontinuities. As a result, the work has mostly been limited to measuring the quality of the mid-vowel concatenations for two Czech speakers—one female and one male.

As the first step, a procedure for collecting perceptual data for the evaluation of the concatenation cost functions have been proposed in this work. The conclusions for this part of the work will be presented in Sec. 6.1. The fundamental role of $F0$ has been demonstrated, more details on the findings will be given in Sec. 6.2. It has also been shown that the impact of phonetic contexts on the quality of the mid-vowel concatenations has been overrated. Some interesting findings have however been discovered for this area as well, and we will conclude on this topic in Sec. 6.3. Finally, we put our work on the concatenation cost functions into a wider scope of the unit selection method as such by proposing a new analytic method for unit selection based TTS systems. Some interesting conclusions of this part of our work will be given in Sec. 6.4.

6.1 Perceptual Data Collection

The first problem that is encountered when working on the concatenation cost functions is the lack of reliably annotated perceptual data.

In order to tackle this problem, the *half-sentence method* has been introduced in our work. The method is very simple and consists in synthesizing sentences by halves. This results in completely natural sentences containing a single concatenation point each. When these sentences are presented to listeners to obtain annotations of the quality of the concatenation points, a high inter-rater agreement can be achieved. This is the advantage of the method over very short stimuli often used in related works. There is also no uncertainty regarding the concatenation point that is rated by the listeners, which is normally the concern for longer stimuli containing multiple concatenations.

To increase the consistency of the annotated data, listener reliability procedures have also been proposed. The procedures can generally be used for different sorts of listening tests.

6.2 Role of F0 Discontinuities

The impact of $F0$ discontinuities on the quality of concatenations is in related works traditionally mitigated by applying different pitch smoothing methods.

This is something we wanted to avoid in our work as any signal modification inevitably introduces a risk of distortions that are difficult to measure and can have an effect on listeners' judgments of the quality of concatenation points.

Our originally planned strategy was to measure the $F0$ discontinuities and to select sentences that would not contain them. For this purpose, we trained SVM classification models using fine-grained pitch contours extracted from neighborhoods of concatenation points as predictors. It has shown that the models can perform very well in identifying the concatenation artifacts in general. This is a very important finding as a widely spread belief has been that the discontinuities have to be measured at least in three different dimensions—$F0$, energy and spectral envelopes.

In fact, using the fine-grained $F0$ contours and calculating their differences by the Euclidean distance is powerful enough for identifying a vast majority of the mid-vowel concatenation artifacts for the two speakers used in our experiments. This could not be proven only for the male voice high vowels. For the female voice vowels /a/,/e/ and /o/, an attention has to be paid to nasal contexts (see more details below). Still, even for the male voice high vowels, the classification models perform very well.

According to our experiments, the $F0$ discontinuities have to be measured, at least for vowels, using pitch contours rather than single values at the concatenation points.

6.3 Role of Phonetic Contexts

As a part of our work, an analysis of the impact of different consonantal contexts on the quality of mid-vowel concatenations has been investigated. To our surprise, no effect has been found for the male voice. For the female voice, the nasal contexts are important for the vowels /a/, /e/ and /o/. It has been shown that for these vowels concatenating diphones from oral and nasal contexts is undesirable as phase mismatches can be introduced at the concatenation points. The phase mismatches appear due to a stronger harmonic component in the context of nasals, which causes misplacements of pitch marks. In such cases, the phase mismatches can be avoided by relabeling the pitch marks. This is however difficult to automate. For the vowels /e/ and /o/, the harmonic component interferes with $F0$ signal peaks, which makes it impossible to avoid concatenation artifacts.

Some minor phonetic context impacts has also been observed for the female vowel /i/. In this case, they seem to be due to energy differences at the concatenation points. This however still needs to be confirmed.

An important finding to mention at this point is that the results of our phonetic analysis are only valid for sentences that do not contain $F0$ discontinuities. It has been shown that different consonantal contexts can affect the variability of the $F0$ contours, which then demonstrates in different discontinuity detection rates for different consonantal groups. This may be an explanation for the difference in our results and the results presented in some other works dealing with the same problem.

6.4 Unit Selection Analytic Method

At least one step has been done in this work towards extending the concatenation cost design to other phonemes than vowels. We have proposed an analytic method that can be used for measuring a perceptual relevance of the unit selection costs and their sub-components.

The method is based on an assumption that the unit selection costs and their sub-components should reflect in their extreme values on the haphazardly appearing audible artifacts in the outputs of the unit selection based TTS systems.

The method was used to analyze the current implementation of the costs and the concatenation cost sub-components in the ARTIC TTS system. It has been shown that a lot of potential exists in improving the costs. For the concatenation cost, this potential has been estimated as 25% in terms of the number of audible artifacts annotated by expert listeners in 50 randomly selected sentences synthesized by an experimental version of the system.

Nevertheless, having an ideal concatenation cost function is still not sufficient for explaining all the audible artifacts as can be seen from Fig. 5.22.

6.5 Contribution of this Thesis

Despite its narrowed scope, we believe that our work has contributed to TTS community in a couple of ways. The key points can be summarized, but are not limited to the following:

- A procedure was proposed for collecting reliably annotated data for the evaluation of the concatenation cost functions. The procedure includes the *half-sentence method* (described in Sec. 2.2) and the listeners reliability analysis (described in Sec. 2.4).
- The fundamental role of fine-grained $F0$ contours in the perception of the mid-vowel concatenation quality has been discovered and demonstrated. It has been shown that the $F0$ contours can be successfully used as predictors for training SVM classifiers for detecting concatenation artifacts.
- Nasals have been for some vowels identified as a consonantal context sensitive to the quality of the mid-vowel concatenations. Phase mismatches that can appear at the concatenation points due to concatenating diphones from oral and nasal contexts have been described.
- An analytic method has been proposed for measuring the perceptual relevance of the unit selection costs. This method can be used for analyzing any unit selection based TTS system. It can also be leveraged for automatic tuning of the unit selection weights.

6.6 Future Work

Many experiments have been presented as parts of this thesis. Still, there is a lot remaining to be done.

One of the most obvious deficiencies of our work is its limitation to the two speakers. This prevents us from generalizing the interesting results that have been presented. An important extension would also be to closely analyze other phonemes and to cover more prosodic contexts.

The described findings should be incorporated into the ARTIC TTS system. This is an ongoing activity, which is supported by using the proposed analytic method. It will be interesting to see how much quality improvement can be achieved for the system by incorporating the $F0$ sub-component of the concatenation cost based on pitch contours or their approximation as a replacement for the static $F0$ differences.

It remains unclear how to measure the discontinuities in the male voice high vowels. The SVM classification models show a good performance, but they are difficult to incorporate into a TTS system. To streamline their incorporation, it is necessary to either collect manually annotated data in different prosodic contexts or to be able to normalize the training data in a way that the models can be better generalized.

Last but not least, the analysis of the current implementation of the unit selection costs has raised an interesting question. One of the hypothesis is that the inability of the unit selection to deliver completely natural outputs free of haphazardly appearing artifacts can be a data scarcity problem. We would like to investigate this hypothesis more closely.

Bibliography

[BDDD04] B. Bozkurt, B. Doval, C. D'Alessandro, and T. Dutoit, "Appropriate windowing for group delay analysis and roots of z–transform of speech signals," in *EUSIPCO '04*, Vienna, Austria, September 2004, pp. 733–736.

[Bel00] J. R. Bellegarda, "Exploiting latent semantic information in statistical language modeling," *Proc. of the IEEE*, vol. 88, no. 8, pp. 1279–1296, 2000.

[Bel04] ——, "A novel discontinuity metric for unit selection text–to–speech synthesis," in *SSW5 '04*, Pittsburgh, PA, June 2004, pp. 133–138.

[Bel05] ——, "Latent semantic mapping," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 70–80, 2005.

[Bel06a] ——, "LSM–based feature extraction for concatenative speech synthesis," in *SAPA '06*, Pittsburgh PA, September 2006, pp. 59–64.

[Bel06b] J. Bellegarda, "Further developments in LSM–based boundary training for unit selection TTS," in *INTERSPEECH '06*, Pittsburgh, PA, September 2006, pp. 1320–1323.

[Bel06c] ——, "LSM-based boundary training for concatenative speech synthesis," in *ICASSP '06*, vol. 1, Toulouse, France, May 2006, pp. 721–724.

[Ben05] C. L. Bennett, "Large scale evaluation of corpus–based synthesizers: Results and lessons from the Blizzard challenge 2005," in *INTERSPEECH '05*, Pittsburgh, PA, USA, 2005, pp. 105–108.

[BHW10] A. Ben-Hur and J. Weston, *A User's Guide to Support Vector Machines.* Springer Berlin / Heidelberg, 2010, ch. 13, pp. 223–239.

[BMR99] M. Beutnagel, M. Mohri, and M. Riley, "Rapid unit selection from a large speech corpus for concatenative speech synthesis," in *EUROSPEECH '99*, vol. 2, Budapest, Hungary, September 1999, pp. 607–610.

[BPQ$^+$99] M. Balestri, A. Paechiotti, S. Quazza, P. Salza, and S. Sandri, "Choose the best to modify the least: a new generation concatenative synthesis system," in *EUROSPEECH '99*, vol. 99, Budapest, Hungary, September 1999, pp. 2291–2294.

[BSF05] I. Bjørkan, T. Svendesen, and S. Farmer, "Comparing spectral distance measures for join cost optimization in concatantive speech synthesis," in *INTERSPEECH '05*, Lisbon, Portugal, September 2005, pp. 2577–2580.

[BT97] A. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *EUROSPEECH '97*, vol. 2, Rhodes, Greece, September 1997, pp. 601–604.

[CBd09] D. Cadic, C. Boidin, and C. d'Alessandro, "Vocalic sandwich, a unit designed for unit selection TTS," in *INTERSPEECH '09*. Brighton, UK: ISCA, September 2009, pp. 2079–2082.

[CC99] J.-D. Chen and N. Campbell, "Objective distance measures for assessing concatantive speech synthesis," in *EUROSPEECH '99*, Budapest, Hungary, September 1999, pp. 611–614.

[CH98] D. Chappell and J. H. L. Hansen, "Spectral smoothing for concatenative speech synthesis," in *ICSLP '98*, Sydney, Australia, 1998, pp. 1935–1938.

[Chi85] L. A. Chistovich, "Central auditory processing of peripheral vowel spectra," *J. Acoust. Soc. Am.*, vol. 77, pp. 789–805, 1985.

[CI97] A. Conkie and S. Isard, "Optimal coupling of diphones," in *Progress in Speech Synthesis*, J. Van Santen, R. Sproat, J. Olive, and J. Hirschberg, Eds. New York: Springer–Verlag, 1997, pp. 293–304.

[CW85] J. Cullum and R. Willoughby, "Real rectangular matrices," in *Lanczos Algorithms for Large Symmetric Eigenvalue Computations.* Boston: Birkhauser, 1985, vol. 1, ch. 5.

[DFC98] W. Ding, K. Fujisawa, and N. Campbell, "Improving speech synthesis of CHATR using a perceptual discontinuity function and constraints of prosodic modification," in *SSW3 '98*, Jenolan Caves House, Blue Mountains, NSW, Australia, November 1998, pp. 191–194.

[DLR77] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[Don01] R. E. Donovan, "A new distance measure for costing spectral discontinuities in concatantive speech synthesizers," in *SSW4 '01*, Perthshire, Scotland, 2001.

[DRO93] V. Digalakis, J. Rohlicek, and M. Ostendorf, "ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 4, pp. 431–442, 1993.

[Dut08] T. Dutoit, "Corpus-based speech synthesis," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Springer Berlin, Heidelberg, 2008, ch. 21, pp. 437–455.

[Faw06] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[FL71] O. Fujimura and J. Lindqvist, "Sweep-tone measurements of vocal tract characteristics," *J. Acoust. Soc. Am.*, vol. 49, pp. 541–558, 1971.

[Fra03] J. Frankel, "Linear dynamic models for automatic speech recognition," Ph.D. dissertation, University of Edinburgh, 2003.

[Fur86] S. Furui, "On the role of spectral transitions for speech perception," *J. Acoust. Soc. Am.*, vol. 80, pp. 1016–1025, 1986.

[GH96] Z. Ghahramani and G. E. Hinton, "Parameter estimation for linear dynamical systems," University of Toronto, Tech. Rep. CRG-TR-96-2, 1996. [Online]. Available: Software at www.gatsby.ucl.ac.uk/ zoubin/software.html

[Han10] Z. Hanzlíček, "Czech HMM-based speech synthesis," in *Proc. of the 13th International Conference TSD 2010, Lecture Notes in Artificial Intelligence*, vol. 6231. Germany: Springer Berlin / Heidelberg, 2010, pp. 291–298.

[HB96] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP '96*, vol. 1, Atlanta, Georgia, May 1996, pp. 373–376.

[HS56] A. S. House and K. N. Stevens, "Analog studies of the nasalization of vowels," *J. Speech Hearing Disorders*, vol. 21, pp. 218–232, 1956.

[HS85] S. Hawkins and K. N. Stevens, "Acoustic and perceptual correlates of the non-nasal–nasal distinction for vowels," *J. Acoust. Soc. Am.*, vol. 77, pp. 1560–1575, 1985.

[HWHK63] A. S. House, C. E. Williams, M. H. L. Hecker, and K. D. Kryter, "Psychoacoustic speech tests: A modified rhyme test," U.S. Air Force Systems Command, Hanscom Field, Electronics Systems Division, Tech. Rep., June 1963.

[Kai90] J. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *ICASSP '90*, vol. 1, April 1990, pp. 381–384.

[KM77] R. D. Kent and F. D. Minifie, “Coarticulation in recent speech production models,” *J. Phonetics*, vol. 5, pp. 115–133, 1977.

[KOS06a] B. Kirkpatrick, D. O’Brien, and R. Scaife, “Feature extraction for spectral continuity measures in concatenative speech synthesis,” in *INTERSPEECH ’06*, Pittsburgh, PA, USA, September 2006.

[KOS06b] ——, “A comparison of spectral continuity measures as a join cost in concatenative speech synthesis,” in *Proc. IET Irish Signals and Systems Conference*, Dublin, Ireland, June 2006, pp. 515–520.

[KOSE07a] B. Kirkpatrick, D. O’Brien, R. Scaife, and A. Errity, “On the role of spectral dynamics in unit selection speech synthesis,” in *INTERSPEECH ’07*, Antwerp, Belgium, August 2007, pp. 2889–2892.

[KOSE07b] ——, “Spectral dynamics as a source of discontinuity in concatenative speech synthesis,” in *Proc. 15th International Conference on Digital Signal Processing*, Cardiff, Wales, UK, July 2007, pp. 615–618.

[KR05] L. Kaufman and P. Rousseeuw, *Finding groups in data: An introduction to cluster analysis*, ser. Wiley series in probability and mathematical statistics. Applied probability and statistics. Wiley, 2005. [Online]. Available: http://books.google.ch/books?id=yS0nAQAAIAAJ

[KT02] H. Kawai and M. Tsuzaki, “Acoustic measures vs. phonetic features as predictors of audible discontinuity in concatenative speech synthesis,” in *ICSLP ’02*, Denver, Colorado, USA, September 2002, pp. 2621–2624.

[KV98] E. Klabbers and R. Veldhuis, “On the reduction of concatenation artefacts in diphone synthesis,” in *ICSLP ’98*, Sydney, Australia, 1998, pp. 1983–1986.

[KV01] ——, “Reducing audible spectral discontinuities,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 39–51, January 2001.

[Lee01] M. Lee, "Perceptual cost functions for unit searching in large corpus–based concatenative text–to–speech," in *EUROSPEECH '01*, Aalborg, Denmark, September 2001, pp. 2227–2230.

[Leg12] M. Legát, "Impact of phonetic context mismatches on quality of vowel concatenations," in *Proceedings of 2012 IEEE 11th International Conference on Signal Processing*, Beijing, China, October 2012, pp. 523–526.

[LK02] K.-S. Lee and S.-R. Kim, "Context–adaptive smoothing for concatenative speech synthesis," *IEEE Signal Processing Letters*, vol. 9, no. 12, pp. 422–425, December 2002.

[LLDW11] H. Lu, Z. Ling, L. Dai, and R. Wang, "Building HMM based unit selection speech synthesis system using synthetic speech naturalness evaluation score," in *ICASSP '11*, Prague, Czech Republic, May 22–27 2011, pp. 5352–5355.

[LLW$^+$10] H. Lu, Z. Ling, S. Wei, L. Dai, and R. Wang, "Automatic error detection for unit selection speech synthesis using log likelihood ratio based SVM classifier," in *INTERSPEECH '10*, Makuhari, Japan, 2010, pp. 162–165.

[LMT11] M. Legát, J. Matoušek, and D. Tihelka, "On the detection of pitch marks using a robust multi-phase algorithm," *Speech Communication*, vol. 53, no. 4, pp. 552–566, April 2011.

[LS12] M. Legát and R. Skarnitzl, "The role of nasal contexts on quality of vowel concatenations," in *Proc. of the 15th International Conference TSD 2012, Lecture Notes in Artificial Intellingence*, vol. 7499. Springer Berlin / Heidelberg, 2012, pp. 551–558.

[LTM12] M. Legát, D. Tihelka, and J. Matoušek, "Is unit selection aware of audible artifacts?" 2012, (pre-print).

[MG03] H. Murthy and V. Gadde, “The modified group delay function and its application to phoneme recognition,” in *ICASSP ’03*, vol. 1, 2003, pp. 68–71.

[MKQ92] P. Maragos, J. Kaiser, and T. Quatieri, “On separating amplitude from frequency modulations using energy operators,” in *ICASSP ’92*, vol. 2, March 1992, pp. 1–4.

[MRTT04] J. Matoušek, J. Romportl, D. Tihelka, and Z. Tychtl, “Recent improvements on ARTIC: Czech text–to–speech system,” in *INTERSPEECH ’04*, Jeju, Korea, October 2004, pp. 1933–1936.

[MTR06] J. Matoušek, D. Tihelka, and J. Romportl, “Current state of Czech text–to–speech system ARTIC,” in *Proc. of the 9th International Conference TSD 2006, Lecture Notes in Artificial Intellingence*, vol. 4188. Springer Berlin / Heidelberg, 2006, pp. 439–446.

[Oli77] J. Olive, “Rule synthesis of speech from dyadic units,” in *ICASSP ’77*, vol. 2, 1977, pp. 568–570.

[PB08] V. Pollet and A. Breen, “Synthesis by generation and concatenation of multiform segments,” in *INTERSPEECH ’08.* Brisbane, Australia: ISCA, September 2008, pp. 1825–1828.

[PS05] Y. Pantazis and Y. Stylianou, “Discontinuity detection in concatenated speech synthesis based on nonlinear speech analysis,” in *INTERSPEECH ’05*, Lisbon, Portugal, September 2005, pp. 2817–2820.

[PS07] ——, “On the detection of discontinuities in concatenative speech synthesis,” in *Progress in Nonlinear Speech Processing.* Springer Berlin / Heidelberg, 2007, vol. 4391, ch. 6, pp. 89–100.

[RG99] S. Rowies and Z. Ghahramani, “A unifying review of linear gaussian models,” *Journal of Neural Computation*, vol. 11, no. 2, pp. 305–345, 1999.

[RM05] J. Romportl and J. Matoušek, “Formal prosodic structures and their application in NLP,” in *Proc. of the 8th International Con-*

ference TSD 2005, Lecture Notes in Artificial Intelligence, vol. 3658. Germany: Springer Berlin / Heidelberg, 2005, pp. 371–378.

[Sag88] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," in *ICASSP '88*, vol. 1, 1988, pp. 679–682.

[SC05] A. Syrdal and A. Conkie, "Perceptually-based data-driven join costs: comparing join types," in *INTERSPEECH '05*, Lisbon, Portugal, September 2005, pp. 2813–2816.

[Sch06] J. Schroeter, "Text–to–speech (TTS) synthesis," in *Circuits, Signals, Speech and Image Processing*, R. C. Dorf, Ed. CRC Press, 2006, ch. 16, pp. 1–13.

[SKN08] S. Sakai, T. Kawahara, and S. Nakamura, "Admissible stopping in Viterbi beam search for unit selection in concatenative speech synthesis," in *ICASSP '08*, Las Vegas, USA, 2008, pp. 4613–4616.

[SS01] Y. Stylianou and A. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," in *ICASSP '01*, vol. 2, Salt Lake City, Utah, 2001, pp. 837–840.

[Swe64] J. A. Swets, *Signal Detection and Recognition by Human Observers: Contemporary Readings*, J. A. Swets, Ed. Wiley, 1964.

[Syr01] A. K. Syrdal, "Phonetic effects on listener detection of vowel concatenation," in *EUROSPEECH '01*, Aalborg, Denmark, September 2001, pp. 979–982.

[Tih05a] D. Tihelka, "Symbolic prosody driven unit selection for highly natural synthetic speech," in *INTERSPEECH '05*, Lisbon, Portugal, 2005, pp. 2525–2528.

[Tih05b] ——, "The unit selection approach in Czech TTS synthesis," Ph.D. dissertation, University of West Bohemia in Pilsen, Department of Cybernetics, 2005.

[TK02] M. Tsuzaki and H. Kawai, “Feature extraction for unit selection in concatenative speech synthesis: Comparison between AIM, LPC, and MFCC,” in *ICSLP ’02*, Denver, Colorado, USA, September 2002, pp. 137–140.

[TKM10] D. Tihelka, J. Kala, and J. Matoušek, “Enhancements of Viterbi search for fast unit selection synthesis,” in *INTERSPEECH ’10*, Makuhari, Japan, 2010, pp. 174–177.

[TM06] D. Tihelka and J. Matoušek, “Unit selection and its relation to symbolic prosody: A new approach,” in *INTERSPEECH ’06*, vol. 1, Bonn, Germany, 2006, pp. 2042–2045.

[Tsu01] M. Tsuzaki, “Feature extraction by auditory modeling for unit selection in concatenative speech synthesis,” in *EUROSPEECH ’01*, Aalborg, Denmark, September 2001, pp. 2223–2226.

[vCR96] H. van den Heuvel, B. Cranen, and T. Rietveld, “Speaker variability in the coarticulation of /a, i, u/,” *Speech Communication*, vol. 18, pp. 113–130, 1996.

[Vep04] J. Vepa, “Join cost for unit selection speech synthesis,” Ph.D. dissertation, University of Edinburgh, 2004.

[Vit67] A. Viterbi, “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm,” *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[VK03] J. Vepa and S. King, “Kalman–based join cost for unit–selection speech synthesis,” in *EUROSPEECH ’03*, Geneva, Switzerland, September 2003, pp. 293–296.

[VK04] ——, “Join cost for unit selection speech synthesis,” in *Speech Synthesis*, A. Alwan and S. Narayanan, Eds. Prentice Hall, 2004.

[VK06] ——, “Subjective evaluation of join cost and smoothing methods for unit selection speech synthesis,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1763–1771, 2006.

[VKT02a] J. Vepa, S. King, and P. Taylor, “New objective distance measures for spectral discontinuities in concatenative speech synthesis,” in *Proc. IEEE Workshop on Speech Synthesis*, 2002, pp. 223–226.

[VKT02b] ——, “Objective distance measures for spectral discontinuities in concatenative speech synthesis,” in *ICSLP ’02*, Denver, Colorado, USA, September 2002, pp. 2605–2608.

[VR93] W. Verhelst and M. Roelands, “An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech,” in *ICASSP ’93*, 1993, pp. 554–557.

[WM98] J. Wouters and M. Macon, “Perceptual evaluation of distance measures for concatenative speech synthesis,” in *ICSLP ’98*, vol. 6, Sydney, Australia, 1998, pp. 2747–2750.

[YS95] B. Yegnanarayana and R. Smits, “A robust method for determining instants of major excitations in voiced speech,” in *ICASSP ’95*, vol. 1, 1995, pp. 776–779.

Résumé

This thesis deals with one of the key aspects of the unit selection speech synthesis method—**design of a concatenation cost function**. The concatenation cost function measures quality of concatenations of units that are taken from a unit database at synthesis runtime. Ideally, the obtained values should correlate well with human perception.

The design of concatenation cost functions is a complex problem due to a wide range of different audible artifacts that can be encountered at concatenation points. Therefore, the scope of the work is narrowed to five short Czech vowels and two speakers—one female and one male. In the first part of this thesis, a method for collecting reliable perceptually annotated data containing a wide range of different concatenation discontinuities is proposed.

It is generally believed that concatenation discontinuities have to be measured at least in three dimensions—energy, $F0$ and spectrum [Dut08]. This work shows that $F0$ is the crucial component for the quality of mid-vowel concatenations. It however has to be measured by using $F0$ contours capturing the dynamics of $F0$ in concatenation areas rather than by calculating static $F0$ differences at concatenation points, which is the traditional approach.

It is shown that different consonantal contexts potentially changing the spectral content of vowels due to coarticulation have only a limited impact.

The work on the concatenation cost design is put into a wider perspective of the unit selection method by proposing an analytic algorithm that allows for measuring the perceptual relevance of different unit selection costs, their sub-components and weights.

Shrnutí

Tato práce se zabývá jednou z klíčových součástí metody syntézy řeči výběrem jednotek—**návrhem ceny řetězení**. Cena řetězení měří kvalitu spojení jednotek, které jsou během syntézy vybírány z databáze jednotek. V ideálním případě by tato cena měla dosahovat vysoké korelace s lidským vnímáním kvality řetězení.

Návrh ceny řetězení je složitou úlohou vzhledem k velkému množství slyšitelných vad, které se mohou v bodech řetězení vyskytovat. Z tohoto důvodu, byl rozsah této práce omezen na pět krátkých českých samohlásek a dva řečníky—jednoho ženského a jednoho mužského. V první části této práce je navržen postup, jehož použitím lze získat velké množství dat se spolehlivým označením kvality bodů řetězení od mnoha posluchačů.

Nespojitosti v bodech řetězení je podle obecného předpokladu nutno měřit ve třech oblastech—energie, $F0$ a spektra [Dut08]. Tato práce ukazuje, že pro kvalitu řetězení v samohláskách je nejdůležitější $F0$. Nespojitosti v $F0$ musí však být měřeny na konturách, které zachycují průběh $F0$ v oblastech řetězení, a ne jako místní rozdíl hodnot $F0$ v bodě řeťězení, což je tradiční přístup.

Je ukázáno, že různé souhláskové kontexty, které mohou měnit spektrální obsah samohlásek, mají pouze omezený vliv na kvalitu jejich řetězení.

Práce na návrhu ceny řetězení je zasazena do širšího kontextu metody syntézy výběrem jednotek. Je navržen analytický postup, který umožňuje měřit percepční důležitost různých cen metody výběru jednotek, stejně tak jejich komponent a vah.

Résumé

Cette thèse porte sur l'un des aspects clés de la synthèse de la parole par sélection d'unités—**la conception d'une fonction de coût de concaténation**. La fonction de coût de concaténation mesure la qualité de la concaténation des unités sélectionnées dans une base d'unités au moment de l'exécution de la synthèse. Idéalement, les valeurs obtenues devraient pouvoir être mises en corrélation avec la perception humaine.

La conception d'une fonction de coût de concaténation est un problème complexe en raison de la grande variété d'artéfacts perceptibles qui peuvent être rencontrés aux points de concaténation. De ce fait, la portée de ce travail a été réduite à cinq voyelles du Tchèque et à deux locuteurs, une femme et un homme. Dans la première partie de cette thèse, nous proposons une méthode pour collecter de manière fiable des données annotées contenant une grande variété de discontinuités de concaténation.

Il est communément admis que les discontinuités de concaténation doivent être mesurées en utilisant au moins trois dimensions : l'énergie, $F0$ et le spectre [Dut08]. Ce travail montre que $F0$ joue un rôle crucial dans la concaténation des voyelles moyennes. $F0$ doit cependant être mesurée en utilisant ses contours qui capturent sa dynamique dans les zones de concaténation, plutôt qu'en calculant les différences de $F0$ statiques aux points de concaténation, ce qui constitue l'approche traditionnelle.

Nous montrons que des contextes consonantiques différents, qui changent potentiellement le contenu spectral des voyelles en raison de la coarticulation, ont seulement un impact limité.

Nous situons ce travail sur la conception d'un coût de concaténation dans une perspective plus large de la méthode par sélection d'unités, en proposant un algorithme analytique qui permet de mesurer la pertinence perceptuelle des coûts de sélection d'unités, leurs sous-composants et leurs poids.