Polyglot Voice Design for Unit Selection Speech Synthesis

Emina Kurtić

Supervisors: Dr. Korin Richmond, Dr. Robert Clark

Master of Science

in

Speech and Language Processing

Theoretical and Applied Linguistics

School of Philosophy, Psychology and Language Sciences

University of Edinburgh

2004

Abstract

Current text-to-speech (TTS) systems are increasingly faced with mixed language textual input. Most TTS systems are designed to allow building synthetic voices for different languages, but each voice is able to ”speak” only one language at a time. In order to synthesize mixed language input, polyglot voices are needed which are able to switch between languages when it is required by textual input. A polyglot voice will typically have one basic language and additionally the ability to synthesize foreign words when these are encountered in the textual input.

Design of polyglot voices for unit selection speech synthesis is still a research question. An inherent problem of unit selection speech synthesis is that the synthesis quality is closely related to the contents of the unit database. Concatenation of units not in the database usually results in bad synthesis quality. At the same time, building the database with good coverage of units results in a prohibitively large database if the intended domain of synthesized text is unlimited. Polyglot databases have an additional problem that not only single language units have to be stored in the database, but also the concatenation points of words from foreign languages have to be accounted for. This exceeds the database size even more, so that it is worth exploring whether database size can be reduced by including only single language units in the database and handling multilingual units on synthesis time.

The present work is concerned with database design for a polyglot unit selection voice. It’s main aim is to examine whether alternative methods for handling multilingual cross-word diphones result in same or better synthesis quality than including these diphones in the database. Three alternative approaches are suggested and model polyglot voices are built to test these methods. The languages included in the synthesizer are Bosnian, English and German. The output quality of the synthesized multilingual word boundary is tested on Bosnian-English and Bosnian-German word pairs in a perceptual experiment.

Acknowledgements

I would like to thank my first supervisor Korin Richmond for invaluable help, advise and support in all matters connected with this project. Thanks also to my second supervisor Rob Clark for many helpful discussions. Thanks Steini for help with recordings and for answering my numerous questions about multisyn scripts at any time of day and night. Thanks also to my other SLP classmates for many nice moments we had together during the course. Many thanks to all participants in my web experiment for their help. Thanks Ed for help with tables. Finally, thanks to my whole family and Ahmet for their unlimited love and support in all things I did along my way.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification except as specified.

(*Emina Kurtić*)

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Text-to-speech conversion (TTS) is necessary for many applications where written input has to be converted into spoken message. These can be simple applications where the machine is required to produce some kind of information for the user, like reading a bank account details, giving various kinds of timetable information or reading cinema programmes. On the other hand TTS is an important part of more elaborated dialogue systems where humans interact with machines. Call center applications, automatic tutoring systems or different kinds of advanced interactive help systems for the blind are some examples. The ability to handle multilingual textual inputs becomes increasingly an important requirement for TTS systems, since more and more applications include elements from more than one language. Apart from foreign proper names, which are traditionally a problem for speech synthesis, the systems also have to be able to handle unrestricted switching between languages in order to synthesize any text given as input.

The present work explores the possibilities of building a polyglot unit selection synthetic voice able to synthesize unrestricted textual input in three languages. The following sections will give an overview over problems and solutions offered so far in concatenative multilingual text-to-speech synthesis. Furthermore, still open research issues will be pointed out which will lead to the outline of the objectives of the present work.

1.1 Speech Synthesis

TTS conversion includes two major processes: linguistic processing including phonetic transcription of the input text and waveform generation. In most TTS systems these tasks are implemented in different modules. The University of Edinburgh's Festival (Black et al. 2002) is an example of a modular TTS system. The output speech is synthesized from the phonetic transcript of the input text and its associated prosodic features in the waveform generation module. The present day text-to-speech systems employ one of the two speech synthesis methods: synthesis by rule or concatenative synthesis.

Synthesis by rule involves applying a set of rules to generate speech sounds from the phonetic transcript of a text with prosodic information. Two types of synthesizers belong to this category: articulatory synthesizers and formant synthesizers. In articulatory synthesizers speech is synthesized from parameters which model the motions of the articulators during production of speech sounds. Formant synthesis involves source-filter model of speech, where the periodic or aperiodic glottal pulse is passed through the filter modelling formant frequencies of the vocal tract. The set of rules for formant synthesis describes how pitch and formant frequencies are changed to produce different sounds. Rules are stored as tables describing lists of parameters for each sound like target formant frequencies, duration of the sound, duration of transitions to the next sound etc. The rules for rule-based synthesizers are to greatest extent manually compiled, although (Holmes & Holmes 2001, ch. 6.5.1) mention attempts to automatize the parameter creation by fitting the rules to the natural speech data.

The concatenative synthesizers produce the waveform by joining and playing back prerecorded units of speech. In this way it is possible to synthesize large number of new utterances from a limited inventory of prerecorded units. The unit size for the prerecorded units can vary from phone and diphone over demisyllable and syllable to whole words or even phrases. It belongs to database design considerations to choose the proper unit size. Generally, larger units mean better quality of synthesized speech but this trades off against size of database which affects search time for the proper units during synthesis and also has practical implications for the database construction as described in detail in chapter 2. Whole word units or larger can be chosen if synthesizer is required to create output from a small domain, known in advance to the system designer. In this case whole word or even larger units can be stored in the database,

so that these cover all possible intended outputs of the synthesizer. This kind of synthesis is called limited domain synthesis and produces generally high quality synthetic speech. However, if the synthesizer is intended for an unrestricted domain, mostly the whole language, it is impossible to store large units for every possible speech event in the database. Thus, in the concatenative synthesis for unrestricted domains sub-word units, most commonly diphones, are used. If only one example of each unit is stored in the database, it will have prosodic features (amplitude, f0 and duration) suitable for the context in which the unit has been recorded. This, however, is not suitable for many other contexts, in which the unit has to be used in the synthesis. Therefore, after unit concatenation, signal processing techniques PSOLA (Moulines & Charpentier 1990), LPC analysis (Hunt et al. 1989) or MBROLA (Dutoit et al. 1996) are applied to modify the prosodic features. However, every signal processing distorts the waveform and affects the quality of the output speech. If the signal processing is kept to the minimum, much of the original voice quality and speaking style can be preserved, so that the resulting voice sounds like the voice of the person whose voice has been recorded.

Concatenative synthesis is currently predominating synthesis method. The main drawback of the rule-based synthesizers is that the synthesized speech sounds rather machine like and lacks in naturalness, compared to the speech produced from recordings of natural speech. This is mainly due to the fact that it is hard to develop rules which capture the full variability of acoustic and prosodic parameters in natural continuous speech. However, concatenative synthesizers too have problems with variability of speech, perhaps with the exception of limited domain synthesizers, where the variability is predefined by the application and can be captured in the database. The quality of a concatenative synthesizer strongly depends on its database. If the unit inventory contains variety of segmental and prosodic contexts, the synthesizer will be able to produce a wider range of good quality utterances. However, it will not be able to cope satisfactorily with any new inputs not covered by the units in the database. Rule-based synthesizers, on the other hand, are much more flexible with regard to synthesis of unrestricted input, both in segmental and prosodic terms. They are adaptable to new segmental and prosodic features since these parameters are easily controlled in the rules.

The flexibility of rule-based synthesizers considering new input makes them theoretically more suitable for synthesis of multilingual speech, when more than one language is used within the same utterance, by the same voice. The voice has to be able to pro-

nounce sounds not in the sound inventory of the basic language of the synthesizer, and it is easy to synthesize any sound by rule although these usually do not sound very naturally. However, it is still a research problem how foreign sounds should be handled by a concatenative synthesizer. Because of the dependency of concatenative synthesizer on the predefined sound inventory and database, the integration of foreign sounds in concatenative synthesizer is a problem of suitable database design.

1.2 Unit Selection Speech Synthesis

Unit selection synthesis (Hunt & Black 1996) is a concatenative synthesis method in which predefined units are selected automatically from a large database of natural speech. It is a data driven approach to speech synthesis, which makes use of increased storage capabilities in computers.

Before unit selection, concatenative synthesis involved concatenation of units (usually diphones) from fixed databases, i.e. databases which contained only one example of each unit. However, having only one example of each unit in the database can not account for variation in pronunciation generally found in natural speech. Segmental co-articulation effects spread, as it is generally known, also across more than one phone or diphone. Additionally, prosodic factors like stress, position within the syllable or intonational phrase affect the pronunciation of a unit. Correct prosody is achieved here by signal processing techniques which distort the waveform and impair the quality of the output. Also high frequency of unit concatenation points proved to affect the quality of speech, since it resulted in more audible joins between the units.

The primary motivation for unit selection synthesis was to improve synthesis quality by reducing spectral mismatches at the points where units are concatenated. This is achieved by storing multiple examples of a unit recorded in different phonetic and prosodic contexts in the database, and choosing the proper unit for the given context, automatically, at synthesis time. Multiple examples of each unit in different contexts should account for segmental and prosodic variation in the pronunciation, so that post-selection signal processing is minimized. The resulting synthesized speech has more natural variation and is minimally distorted by signal processing techniques. As in fixed databases for concatenative synthesis, the unit size in unit selection databases can be set to phones, diphones or larger units. Also mixed sized units are possible.

The units are selected for synthesis if they minimize the sum of join and target costs (Campbell & Black 1995, Hunt & Black 1996). Join costs measure how well the selected units concatenate. The join costs of units recorded together are zero. Another factor to be considered is target costs. These reflect how well the candidate unit from the data base matches the target unit which is an ideal unit for a given context. Target costs are associated with number of features like position in the utterance, stress, syllable position, F0 shape etc. While the join costs are relatively straightforward to calculate from the waveform, target costs are complicated because they involve both continuous and discrete cost factors. It is also not straightforward to determine how much weight should be assigned to single features. Continuous values are prosodic features like F0, duration or energy. Discrete values are stress, position in syllable, word or phrase, phonetic environment etc. One way to deal with target costs is to encode assignment of costs to different cost factors in rules, which are typically hand-written. A desirable solution however, is to determine target costs automatically. The latter approach to determining target costs is implemented in Festvox, a voice building toolkit (A. & Lenzo 2000) which is used in this project for building the voices.

1.3 Multilingual vs. Polyglot Speech Synthesis

Including new languages into text-to-speech synthesis systems is interesting and useful both for commercial applications and research. Most of the common commercial applications, like reading cinema programmes or telephone book entries, require speech synthesis systems, which are able to synthesize foreign names, foreign street names or names of the movies in the original form. Thus, these rather simple applications already require systems able to handle phones from several different languages. From the research point of view the extendability of the synthesizer to new languages is a challenge. The main concern here is to develop more general systems which would be easily adaptable to new languages. This requires general, language independent algorithms and system architectures. The voice itself, once built for a language can be used for research on that particular language.

Most existing TTS systems are multilingual, in the sense that they allow voices in new languages to be built more or less easily. In an ideal multilingual system the language specific information would be completely separated from the algorithms. The algorithms should be shared across languages, so that only language specific components

of the system have to be changed for a new language. The existing TTS systems handle this in different ways and most of them can successfully integrate new languages. Thus, in a multilingual system like Festival, Bell Labs TTS and many commercial TTS systems there will typically be voices for several languages, but all voices will ”speak” only one language at a time, i.e. they won’t be able to include foreign pronunciations in the synthesis of a single language.

A polyglot voice, on the contrary, should be able to switch between languages if this is required by the textual input to TTS. Such a voice should be able to ”speak” more than one language simultaneously, comparable to the polyglot human speaker, who can switch between the languages if necessary. Adapting a multilingual concatenative TTS system to a polyglot one is still a research question. In the next section, several suggestions made so far on this way are presented. This project is concerned with particular problem of polyglot voices, which is integration of foreign words in a native language sentence.

1.4 Previous Work on Polyglot Speech Synthesis

The approaches to the polyglot speech synthesis, suggested so far, can be grouped into two main groups, according to the way foreign sounds are integrated into the basic language inventory.

The first way of dealing with foreign sounds is to expand the inventory of the basic language of the synthesizer by integrating foreign sounds into it. This approach has been explored by (Traber et al. 1999). This work focuses on an automatic procedure for extracting diphones for four languages from recorded nonsense words. The result is a multilingual diphone inventory for polyglot diphone speech synthesis. The basic language of the system is German but inclusion of Italian, French and English at synthesis time is possible.

Description of the unit selection database for Bell Labs German TTS system (Möbius et al. 1997) also mentions extension of the German diphone database by English interdental fricatives and glide /w/ and French nasalized vowels. This extended inventory should account for foreign phones commonly occurring in foreign words and names.

Eklund & Lindström (1998) base their decision to extend the Swedish phoneset by adding foreign (English) phones on their speech production studies on Swedish sub-

jects. In these studies, 491 Swedish subjects chosen across different age, gender, educational level and native regions were asked to produce sentences containing English words and names. Results are reported in (Eklund & Lindström 1996) and (Eklund & Lindström 1998). The observations are made along two dimensions. One of them is how aware speakers are that the sound to be produced is not Swedish. The other dimension is how well the speaker manages to produce the foreign sound. The results show that Swedish speakers are mostly aware of difference in English pronunciations and extend their sound inventory when pronouncing words of English origin. However, the results of the study might only be valid for Swedish subjects. There are many linguistic and non-linguistic factors, which influence the pronunciation of foreign words, and the country the speaker and listener come from might be one of them. Due to the lack of studies for languages other than Swedish it is difficult to make any generalizations. Eklund & Lindström (1998) also report on integration of English sounds in a Swedish TTS system. However, only preliminary informal evaluations are reported suggesting that including foreign sounds in TTS outputs better quality synthesis than using only Swedish phoneset.

Second type of approach to handling foreign sounds involves replacing them by the closest matching sound from the basic language.

Badino et al. (2004) report on an algorithm for automatic determination of similarity between foreign sounds and sounds of a basic language of the synthesizer. In order to compute the similarity between the sounds, first, phonemes are represented as vectors of articulatory features. Then, the weight of single features in the similarity estimate is determined. Finally, the degree of similarity between the features is calculated. In order to determine perceptually valid weights of single features, an iterative method has been applied, where initially set weights are re-estimated in accordance with native speaker judgements of similarity between the sounds. This approach is based on solution to mapping between English and Japanese in CHATR TTS system previously proposed by (Campbell 2001). Cambell's approach also includes finding the closest equivalent in the native language database based on similarity between articulatory features and using it for synthesis of foreign words. However, unlike in the approach by (Badino et al. 2004), the closest matching sound is not defined by perceptual weighting of articulatory features, but by computing acoustic and prosodic similarity to the model pronunciation synthesized with a native speaker voice.

1.5 Problems with Current Approaches to Polyglot Synthesis

1.5.1 Coverage Problems

Each of the approaches described above has drawbacks. The main disadvantage of extension of the database by foreign phonemes is that it becomes more difficult to find a good compromise between database size and unit coverage. As previously shown in (van Santen 1997, Saikachi 2003, Bozkurt et al. 2003), the diphone types have a long-tailed Zipf distribution with large number of diphone types with low frequencies and only few diphone types with high frequencies. This makes it impossible to provide enough examples of diphones even in a single language database. Adding foreign units means extending the unit inventory to cover which adds to the coverage problem. Analyses of unit distributions in chapter 2 will point out the unit coverage problem in polyglot databases more clearly.

1.5.2 Units not in the Inventory of the Basic Language

Having only basic language units in the database and approximating the foreign ones by these is only appropriate for languages with very similar phone inventory. The problem with this approach arises in cases where there is no one to one matching between a unit in basic language language (L1) and the units with similar acoustic features in language 2 (L2). Three cases can be distinguished here.

First, an acoustically similar unit might not be found in the L1 inventory. German vowel space for example is much larger than Bosnian. The front close-mid rounded vowel /ö/ is not in the vowel inventory of Bosnian. If the German word ”könnte” for example is to be pronounced by a Bosnian voice, the closest match considering all features but roundness is the unrounded close-mid vowel /e/. This, infact, is very common nativization of the German phone by Bosnian native speakers. However, Badino et al. (2004) mention that round/non-round differentiation strongly affects the perception of similarity, so it might be that the algorithm would prefer to neglect differentiation in frontness and choose /o/ instead. In the first case the produced word would be unacceptable and in the latter case the substitution would render syntactically inappropriate word.

Another case of mismatching is than L1 has one unit which is acoustically similar to two or more units in the L2. Bosnian for example has larger affricate inventory than English. For the English affricate /ch/ as in ”chalk” there are at least two similar affricates in Bosnian. One is /tS/ as in the word ”čar” (charm) which is slightly less palatalized than the English phone with same IPA transcription. The other variant is /tc/, the alveolo-palatal fricative as in ”ćar” (profit). The two Bosnian affricate contrast word initially but the contrast would disappear when they are pronounced by an English voice. Replacing the two fricatives by the closest English equivalent would render unclear ambiguous pronunciation.

Finally, further problems can arise if L1 does not have lexical prosodic features and L2 does. Bosnian for example is a word-accent language similar to Swedish (Remijsen & van Heuven 2004). Four different accents can be distinguished. Usage of wrong accent of a word renders grammatically incorrect utterance. In the sentence ”Dosta mi je ovih žena” (*I am fed up with these women*) the genitive plural version of ”žena” (*women* genitive plural) bares long raising accent (cf. section 2.4.1, chapter 2). Another word with similar articulatory but different prosodic features is nominative singular žena (*women* nominative singular). This would use short raising accent which would not agree in case and numerus with the demonstrative pronoun ”ovih”. Thus, for word-accent languages like Swedish or Bosnian, tone languages like Chinese, and lexical stress languages like Japanese, it is insufficient to rely only on feature vectors describing articulatory positions to define similarity between units. Further prosodic features have to be included in addition to articulatory features.

1.5.3 Multilingual Cross-Word Units

Cross-word combinations are problematic even for single language databases since the units at the word boundaries violate phonotactic constraints which normally hold in a language, thus increasing the number od units to cover. For the unlimited domain synthesis, a number of new word combinations not recorded in the database can occur in the textual input. Most databases contain cross-word units for a language since not covering these can result in selections leading to bad quality synthesis or unintelligibility of the output. Unit selection databases require multiple examples of these units in order to selects the best units for given contexts. Considering only optimal coverage of single language units in the unit selection polyglot database will not ac-

count for units at the word boundaries of two words from different languages. Spectral distortions at the concatenation points of words from two different languages and bad synthesis quality can be expected as a result. On the other hand, if cross-word units are added to the polyglot unit inventory and included into the database the compromise between database size and unit coverage becomes even bigger problem than for single language databases. Thus, finding a satisfactory way of dealing with cross-word units is an important issue in building polyglot databases.

1.5.4 How Native Should a Polyglot Voice Sound?

Different method of handling foreign sounds in a polyglot speech synthesis system are closely connected to the question how close to the pronunciation of a native speaker of the target foreign language the foreign pronunciations should be.

The approach by (Campbell 2001, Badino, Barolo & Quazza 2004), where the sounds not in the sound inventory of the basic language are replaced by the perceptually closest matching sound of the native language results in completely nativized foreign sounds. It can be argued, in favour of this approach, that it is the way of human multilingual speech production. It is the fact that not many polyglot human speakers will speak all languages they are familiar with without foreign accent. They will rather nativize foreign sounds to varying extent to the pronunciations of their native language. Even when a speaker is aware of foreign pronunciation, he might not employ it. Various linguistic and socio-cultural factors influence the extent of nativization of the foreign sounds. Eklund & Lindström (1996) mention ”speaker’s competence and performance capabilities with respect to the source language, the speaker’s expectations of the listener’s competence, the relative social status of speaker and listener, the socio-cultural distance to the country of origin, recency and frequency of the lexical item in question and similarities/dissimilarities between the two phonological systems in question” as some of the influencing factors. Eklund & Lindström (1999) investigate the influence of age, gender and dialectal origin on nativization of English sounds in Swedish and come to conclusion that age is a significant factor influencing the extent to which the foreign sounds are nativized in the production. Additionally, phonetic considerations like co-articulation effects of the basic language and economy of effort in producing foreign sounds may play a role too. In spite of nativization, the non-native speech is intelligible and mostly accepted by the native speakers of a language. If this is so, than

it could be claimed that the sound inventory of the native language (or a basic language of the TTS system) is sufficient to cover both native and foreign pronunciations.

Another possibility is to adopt the approach of (Eklund & Lindström 1998, Traber et al. 1999, Möbius et al. 1997), which means to expand the sound inventory of a language by foreign sounds and come closer to the native foreign pronunciations. This strategy also seems to be consistent with multilingual human speech production as the production studies on Swedish by Eklund and Lindström suggest. These studies would support inclusion of foreign sounds into the sound inventory of a language since this would possibly reflect the common way humans deal with foreign sounds and thus improve the quality of the synthesized speech. However, as also noted in (Eklund & Lindström 2000), it is not clear to what extent the foreign sounds should be included. Minimizing the foreign sound inventory would lead to higher nativization, whereas maximizing it would lead to perfect pronunciation of the foreign words. The latter would not be typical for humans any more, and the question is whether it would be acceptable by human listeners. Thus, the studies do not offer the answer to the question how native the synthesized speech should be. Furthermore, expanding the sound inventory also introduces many practical problems for concatenative synthesis as it will be discussed later in more detail. One of them is the choice of the speaker for recording the database. Whereas it is relatively easy to find a bilingual native speaker, the task is more complicated, and even impossible, if four or more languages should be synthesized, or if any arbitrary language combination is required.

How native a polyglot voice should be can in the last instance only be decided by extensive perception and acceptability studies, or more practically, by the requirements of the application. In the present reports on both including and not-including foreign sounds in the inventory of a language only informal evaluations are described. Thus, although expanding the inventory by foreign phonemes seems to be close to the human production mechanisms, as studies on Swedish suggest, it is not clear whether it yields better quality speech than replacing foreign sounds by perceptually similar native ones, when the quality is measured as subjective acceptability. Given these facts, how native the polyglot voice should be was not a concern of this project. The quality of the synthesis is judged only by the spectral quality of the output sound. The voice built in this project has limited nativeness when pronouncing English and German words due to the choice of the speaker. For practical reasons my own voice was recorded and the database of units was constructed from these recordings. Thus, as the synthetic unit

selection voice sounds like the recorded voice, the voice built here will have foreign accent in English and German. The focus in the project however, is on finding methods for language switching, which can be applied generally in building polyglot voices, and getting a more native voice is only the question of having a more native speaker to record.

1.6 Objectives and Outline of the Thesis

The wider objective of this project was first to build a polyglot unit selection voice with Bosnian as basic language, but able to switch to English and German in any arbitrary context if this is required by the text input. Several decisions had to be made on this way. As outlined in section 1.4 approaches to building polyglot voices differ in the way of organizing the database for polyglot voices and each approach is problematic. Thus the first decision to be made was which approach to the database design to adopt or how to combine the approaches to minimize their disadvantages. Initial investigations of unit distributions showed that dealing with cross-word units when the words are from different languages is particularly problematic for finding a compromise between coverage and size in the design of polyglot databases. The project thus focuses on finding a way of reducing the number of inter-language cross-word units in polyglot databases and finding alternative ways of dealing with these units in the synthesis. To do this, first possibilities of reasonable coverage of cross-word units in a polyglot unit selection database for unlimited domain have been theoretically examined. The results of these analyses are described in chapter 2. Since the results suggest that satisfactory coverage of units from all three languages is impossible, alternative ways of dealing with cross-word units are explored. These are discussed in chapter 3. Four voices are built from databases implementing these approaches. Finally, the different methods for handling cross-word units are compared experimentally, by quality judgments of output speech synthesized from different databases. Chapter 4 describes experimental goals and design, as well as material used in the experiment. Chapter 4.2.2 describes general voice building procedures and chapter 4.2.3 the voices built for the experiment and synthesis. The results of the experimental assessment of different database designs are presented in chapter 4.3.

Chapter 2

Corpus Analysis

In this chapter construction of a polyglot database containing units from three languages, Bosnian, German and English is discussed. The optimal database contains at least one example of each unit in different predefined contexts. Since such a database is prohibitively large even for a single language, alternative possibilities of finding a compromise between unit coverage and database size in a polyglot database are explored. Unit size is set to diphones. Investigation of different unit sizes show that diphone is the best unit size for a Bosnian, German, English database. The distributions of diphones in three single language corpora is analyzed, and possibilities of creating a single polyglot database out of single language sources is examined. It is shown that although some frequency weighted diphone coverage for single language units can be achieved in a polyglot database, the inclusion of multilingual cross-word diphones (i.e. diphones at the word boundaries of words from different languages) expands the database substantially, so that a creation of such a database is prohibitive.

2.1 Unit Coverage

The optimal database for synthesis of text from certain domain should cover every unit which can possibly occur in the speech in different acoustic and prosodic contexts. For unrestricted domain there is a large number of different units and contexts to be covered. For example, if position in the word is a context parameter with three values, word-initial, word-medial and word-final and the unit is diphone this would already mean that the there have to be at least three examples of each diphone for each con-

text in the database. For English which typically has 1600 diphones the number of diphones would triple to 4800. In order to account for natural variation further context parameters with two or more values are needed. Some possible contexts are stress, position relative to the syllable boundary, position within the phrase, surrounding phones, etc. Thus databases for unrestricted unit selection synthesis easily become very large.

Large databases are constructed from large text corpora by reducing the whole corpus to a subset of sentences which are representative in terms of unit coverage for the whole corpus and then recording a speaker reading these sentences. The main problem with having large databases is time and human power required for recording them. The database should not only be optimal in terms of coverage of units but also in terms of quality of the recorded speech. The speaker should be able to speak clearly and consistently throughout the recordings. There are natural limitations to the ability of a speaker to speak consistently over long periods of time. Also total time needed for recording the database is limited to some reasonable recording time. In addition to these practical matters large databases also require longer search time in the automatic search for best units at synthesis time. Pruning techniques can be applied to reduce the search space, however there is always a possibility to prune the optimal unit and choose an inappropriate one instead, which has a direct impact on the output speech. A further issue in having large databases is their annotation. Accurate annotation always requires manual correction of automatically labelled database. Campbell & Black (1995) mention that accurately annotated smaller database renders better synthesis than purely automatically annotated large database. Thus database design always means finding a compromise between optimal coverage and database size.

The main question to address is how large the database should be, so that the optimal coverage is achieved. A definition of optimal coverage is suggested by (van Santen 1997). van Santen (1997) defines the *coverage index* of a given database with respect to a domain as the probability that all units occurring in a randomly selected test sentence are present in the database. The units used in the study are diphones containing contextual information on accent (accented vs. unaccented) and position within utterance (initial, medial, final) represented in a vector for each diphone. van Santen's results suggest that no reasonably sized database can have optimal coverage for unrestricted domain, even when only two context parameters are considered. The database of 25,000 units had coverage of 0.03, i.e. the probability that all units of a sentence are in the database is only 3%. Coverage index of 0.75 would require at least 150,000

combinations which already is prohibitive in terms of recording time. Coverage also decreases if the text genre used for database construction differs from the genre of the test sentence set (van Santen 1997, Bozkurt et al. 2003).

Since it is impossible to attain an optimal unit coverage in a database for unrestricted domain, several attempts have been made in approximation of the optimal coverage. One suggested solution is to cover most frequently occurring units and discard the rare ones (François & Boëffard 2001, Saikachi 2003). It is based on the fact that covering more frequent units renders higher overall coverage (François & Boëffard 2001) and on the assumption that if less frequent units are not synthesized well, the perceived quality of speech will not be substantially impaired (Campbell & Black 1996). François & Boëffard (2001) use triphonemes (sequences of three phonemes) as units and a mixed genre corpus. They report that removing triphoneme types with less than 10 tokens results in keeping 70% of distinct types and overall coverage of 99.9% for all triphoneme tokens in the corpus. To approximate optimal coverage, they remove all rare triphoneme tokens and include 10 tokens of types occurring more than 10 times. However, relying entirely on covering most frequent units might not be the best solution for every database for two reasons. First, frequency counts are based on single corpora and can not be generally transferred to any random test set especially across text genres. Bozkurt et al. (2003) among others show that coverage of triphone units is best for the corpus the database sentences are selected from and is substantially lower for other corpora. The other reason for not leaving out rare units out of the databases is that they are common in speech. The probability that a rare unit occurs in a random test sentence almost approaches certainty. It is a common distribution of language events that few units have large number of tokens and a very large number of units occurs very rarely. This phenomenon is called ”Large Number of Rare Events (LNRE)” (van Santen 1997). Beutnagel & Conkie (1999) report that rare units are preferred in automatic selection of the units from the database and that inclusion of rare units in the database results in better quality synthetic speech.

In the polyglot database foreign units are integrated in the unit inventory of a basic language of the synthesizer. The number of units in the database can be reduced if some units are shared between languages. However, it is a question to which extent sharing is possible or desired for the nativeness of the voice if this is required for the application. Sharing of units was not investigated in this project, so it is assumed that no units can be shared between languages. In any case, the database has to be extended

to include examples of the foreign units and the combinations of the basic language units with the foreign ones. It follows from this that the trade off between size of the database and unit coverage becomes even more problematic than in the monolingual database.

Covering cross-word units (i.e. units across word boundaries) leads to extension of the database even in a single language case because phonotactic constraints which restrict the number of sub word unit combinations within words do not hold across word boundaries. If the foreign units are added the number of unit combinations at the word boundaries increases. This increase goes along with the increase in LNRE since it can be expected that many of the native-foreign unit combinations at the word boundaries will not occur very frequently.

Hence, there seems not to be an optimal or a generally satisfactory approximate solution for the unit coverage in a open domain unit selection database even for a single language. For open domain synthesis it is probably only feasible to cover the most frequent units in several contexts. Rare units can be handled either by including them too into the database to certain extent or by having a trained rule system in the synthesizer, e.g. a decision tree able to handle unseen events by generalization from the trained cases. It can be expected that good coverage becomes even more problematic in polyglot databases where units from more than one language are covered. Frequency analyses described in the following sections will illustrate the coverage problems in building a polyglot database for unrestricted domain.

2.2 Corpora

As mentioned above the unit coverage of the database relative to the intended output domain depends on the genre of the text used in the creation of the database (van Santen 1997, Bozkurt et al. 2003). The coverage will typically be better if the input text to the synthesizer is from the same genre as the text used for the database. If the domain is unrestricted, i.e if it should be possible to synthesize any sentence of a given language, it is not straightforward to define the type of text that should be recorded. In the synthesis of unrestricted polyglot text, where foreign words are included into the native language text, there is the additional problem that it can not be easily determined when a switch between the languages will occur. Thus, in addition to the problem of cov-

ering the genre there is the problem of finding a single multilingual corpus containing enough code switches to cover all possible unit combinations which can be required to synthesize an arbitrary input to the synthesizer.

In order to get all possible cross-word diphones in the three languages, three single language corpora are used for analysis. The corpora for Bosnian, English and German are compiled from texts downloaded from the internet. In each corpus two genres are covered: literary texts, philosophical texts and newspaper articles. The genre coverage is not optimal for unrestricted domain. However, it is sufficient for analysis purposes presented in the remainder of the chapter. The statistics about the three corpora are given in table 2.1.

Corpus	*Number of Words*	*Number of Sentences*	*Number of Phrases*
Bosnian	572,031	30,768	104,969
English	2,255,293	62,684	245,892
German	1,337,282	50,604	120,721

Table 2.1: Corpora Statistics

English and German corpus were transcribed (phonetized) using Festival synthesizer's front end. For Bosnian corpus a set of letter-to-sound rules was written in Perl. For English transcription the American English phoneset "radio" was used. German Festival uses reduced German celex phoneset. For Bosnian a phoneset was defined. Grapheme-phoneme correspondence is very high in Bosnian, so the phoneset corresponds to the alphabet, additionally including silence phones. It is stated in the literature (Brabec et al. 1952) that there is always a syllable boundary between two vowels in Bosnian, i.e. that diphtongs do not exist. Following this vowel combinations are not included in the phoneset. Festival uses pronunciation dictionaries for transcription of German and English. No available pronunciation dictionaries for Bosnian could be found, so a set of hand written letter-to-sound (LTS) rules was used instead. Due to high grapheme to phoneme correspondence, letter-to-sound rules can be hand written rather than learned from the data. Transcription using pronunciation lexicon has the advantage that additional information about stress, syllable and word boundary is provided, whereas no such context information is available for the text phonetically transcribed only by LTS rules. Thus for Bosnian corpus additional syllabification and stress assignment rules had to be implemented. The syllabification rules were implemented in Perl using the rules for determining syllable boundary indicated in (Brabec et al. 1952) and

(Šipka, personal communication). Implementing stress assignment for Bosnian was not straightforward (cf. section 2.4.1), so it was left out.

Phrase breaks are determined by Festival's module *Phrasify*. For phrase breaks prediction a probabilistic model is used (Black et al. 2002, chapter 17). This model predicts phrasing of an utterance using the probability of a break after certain words, based on their part-of-speech and a general distribution of phrase breaks. Viterbi decoder is used to find the optimal phrase breaks for an utterance.The number of phrases is higher than the number of sentences since every sentence boundary is also a phrase boundary. For Bosnian corpus, phrases determination is based on punctuation. This is the simplest phrasifying method which does not give very good results. However, applying more elaborated phrase prediction methods would require at least more accurate tokenization and tagging of the Bosnian corpus. Since no resources like lexicon or tagged corpora were available for Bosnian, this was out of scope of this project.

After phonetisation of the corpora diphone frequency distribution analyses were made in order to define good coverage for a polyglot database containing units from all three languages. Both context independent and context-dependent diphones are considered in the analyses. Finally, the necessary inclusion of cross-word diphones in this three-language database is discussed. It is shown that this renders a prohibitively large database for all three languages.

2.3 Unit Size

The unit types commonly used in speech synthesizers are phones, diphones, triphones, syllables, demisyllables, words and phrases. The newest release of Festival speech synthesizer, Festival 2 is based on diphones (Clark et al. 2004). Since the intention was to use this synthesizer in this project, the unit size was set to diphones. However, analyses below also show, that diphone is a reasonable unit size if the relationship between database size and unit coverage is considered.

It is known that larger units generally produce better quality synthesis. However, this trades off against larger size of the database because the larger the units, the more units in the database are needed to attain good unit coverage. Table 2.2 gives distinct type counts for different unit sizes as they occur in Bosnian, English and German corpus respectively. The number of distinct unit types for units larger than phones is

Unit size	*Unit type count*			*Units occurring less than 10 times*		
	Bosnian	*English*	*German*	*Bosnian*	*English*	*German*
phone	31	45	47	0	0	0
diphone	907	1,517	1,965	164	146	330
syllable	13,034	11,285	12,226	9,441	6,177	7,638
triphone	24,659	27,314	27,193	14,822	10,479	12,700
word	43,725	37,096	59,739	40,663	29,064	53,297
phrase	71,048	223,418	114,059	70,940	222,931	113,969
sentence	28,636	59,422	49,178	28,601	43,949	48,887

Table 2.2: Type counts of context independent units for different unit sizes

generally lower than theoretically possible number of combinations. This is of course first due to the limitations of the corpora, which never can cover all possible units occurring in the language. However, in addition to this, the space of really occurring units (except phones) is also restricted by phonotactics of the language which exclude certain combination of units. Table 2.2 shows that the number of distinct unit types increases with increasing unit size. Phrases and sentences do not follow this since each sentence end is also a phrase end, so there are more phrases than sentences in total and also more types. For the database construction this means that the larger unit size, the larger number of units is needed in order to attain same coverage of a domain. For example, the number of distinct types for sentences and phrases is about 95% of the total number of sentences and phrases given in the second column of the table 2.1. This distribution means that only few phrases or sentences occur more than once in all corpora. The sentences with occurrence higher than 1 are mostly headings from chronicles in newspapers and single word sentences. There is higher number of phrases with frequency higher than one which are not only one word phrases. However, the most frequent phrase in English for example is ”Oh”, followed by ”He said”. Hence, phrases and sentences with higher frequency are too short to provide good coverage of any domain. This implies that taking units larger than word is unsuitable for a unlimited domain even for a single language. If an arbitrary test sentence is input to the synthesizer, it is almost certain that it will not be covered by the units in the database.

Frequency distributions of sub-word units for German corpus are given in Figure 2.1.

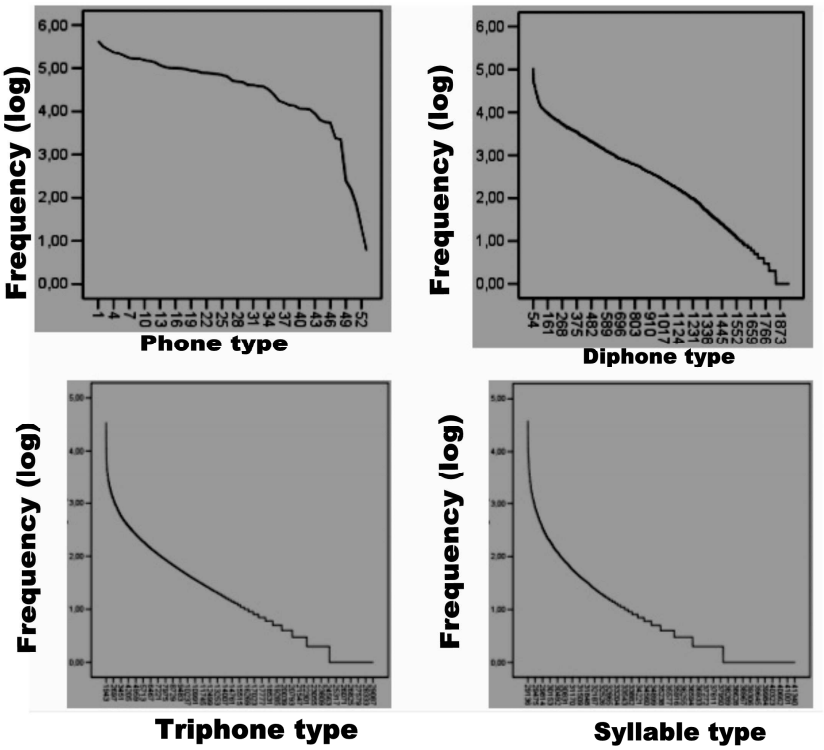

Figure 2.1: Frequency distribution of German sub-word units

German corpus is selected as an example, but in fact all three corpora exhibit similar distribution of units. This means that the larger the unit, the less units with high frequency exist. In terms of domain coverage, these distributions present a problem. The LNRE problem described above becomes more significant for larger units. The larger the unit the more rare units occur and the probability that a unit not covered in the database occurs in an arbitrary test sentence increases. Already for word sized units in Bosnian corpus for example only 7% units occur more than 10 times as indicated in table 2.2. Keeping in mind that the final goal is to construct a database covering enough units for all three languages, words are almost certainly not suitable units in terms of database size needed for acceptable coverage of units.

It has been shown that spectral distortions occur across syllable boundaries rather than within syllables (Yi & Glass 1998), so syllable might be an appropriate unit from the quality point of view. Definition of syllable is not always clear. In the statistics presented above syllable structure for English and German was built by Festival using syllable structure indicated in the dictionaries. For Bosnian syllabification rules have been implemented as mentioned above. Kishore & Black (2003) show that syllable based synthesizer for Hindi performs better than other units. They also note however that this is due to the regular syllable structure and in Hindi. Some units are obviously more appropriate for certain languages, depending on the phonological structure of the language. Since neither of the three languages has regular syllable structure, syllables might not be the best type of unit to cover for the given languages. Also in terms of number of units which have to be covered, syllables do not seem to be a good choice for a polyglot database. The LNRE problem with syllable-sized units is substantial. In German corpus for example 37.52% of syllables occurs more than 10 times. Thus syllable coverage also requires large number of units.

These facts show that diphone is a reasonable unit size. Diphones store transitions between single phones and avoid concatenating single phones in unsteady regions. However, diphone units are not optimal. First, the problem of spectral discontinuities resulting in audible joins is not solved. Spectral mismatches can also occur in the stationary parts of the phones, not only at phone transition points. Thus diphone concatenation points can also sound bad if two diphones originally recorded in different contexts are excised for the synthesis. Choosing diphones as a unit assumes that co-articulation phenomena only spread over at most two phones, which is not true in general. These problems are reduced to some extent in unit selection synthesis, where the best context is found automatically. Units longer than diphones can be chosen if their target and join costs are low. Low join costs mean that spectra of the diphones fit together well, which accounts for some co-articulation phenomena spreading over more than diphones. However, in unit selection it is assumed that diphones in enough different contexts are provided in the database, so that the best ones can be chosen.

Although diphones are not optimal units to be covered in the database, they are still widely used. Relatively small number of diphones is needed for complete context independent coverage of single language databases. Coverage of all diphones in the three languages database would require 4,389 diphones in total, according to the diphone type counts given in table 2.2. Since unit selection requires covering of more than only

one context independent diphone, further possibilities of covering different contexts are explored and described in following section.

2.4 Frequency Distribution of context-dependent diphones

For unit selection diphones in different acoustic and prosodic contexts are required. There is a number of contextual features which influence realization of a diphone. Here, the following, merely prosodic, features are discussed: stress, position relative to syllable boundary, position relative to word boundary and position within the phrase.

2.4.1 Stress

Stressed syllables differ from unstressed ones in pitch, duration and intensity, but some studies (Sluijter & van Heuven 1996) suggest that duration is the main acoustic correlate of stress. This means that units in stressed syllables will have different acoustic properties than same units in unstressed syllables and should thus be distinguished.

Stress information for English and German words is derived from pronunciation lexicons where syllables are marked as either stressed or unstressed. Stress as context parameter has two values which is a simplification. Further distinction between primary, secondary and tertiary stress could be made if pronunciation lexica contained necessary information.

Bosnian has more complex word prosodic system. It is not lexical stress language like English and German but is often characterized as word-accent language (Remijsen & van Heuven 2004). Word accent is a combination of vowel length and pitch contour on the vowel which is encoded in the lexicon. There are four word accents. These are shown in table 2.3 with their traditional notations.

		Pitch	
		raising	falling
Vowel	long	á	â
length	short	à	ä

Table 2.3: Word accents in Bosnian

Stress type	*Example*	*Type count*	
		English	*German*
1	A'B	834	802
2	AB'	1263	1284
3	AB	1275	1709
4	A'B'	303	186
Total		3,675	3,981

Table 2.4: Frequencies of stress context variations for diphones. ' indicates the position of stress in a model diphone AB

The four accents are flexible and thus not easily predictable. However, there are some rules for their distribution (Ivić 1958). Falling accents for example, occur almost only on first syllable and can also occur in monosyllabic words. The raising accents occur on all syllables except the last which prevents them from occurring in monosyllabic words.

Orthographically identical words, e.g. inflectional variants of a noun, can be distinguished by varying word accent. An example of minimal pair for long and short falling accent is "grȃd" (city) and "grädˋ" (hail). Raising accents contrast for example in "žéna" (woman, genitive, plural) and "žèna" (woman, nominative, singular).

Since a pronunciation lexicon could not be found for Bosnian the information on word accent was not available. If a lexicon was available, LTS rules could be trained from it, and the word accent could be predicted for words not in the lexicon. Without a pronunciation lexicon word accent of Bosnian words could not be determined and was not considered in the statistics. Thus the diphone type counts for stress-dependant diphones were done only for English and German as presented in table 2.4. The statistics below show that already for two languages the number of stress dependant diphones increases substantially compared to context independent diphones.

There are four possible differentiations between diphones according to the placement of stress. These are given in table 2.4. Four different context variations means that context stress has four different features. So considering stress as context will theoretically lead to an increase of the database of approximately 4 times. However, not all diphones occur in all contexts and not all contexts have same frequency. The stress type

frequencies in the table indicate that diphone types with variation 3 where both phones are unstressed are more common than diphone types with other variations. Type 3 diphones are followed by the stress type 2 where the stress is on the second phone. The fourth diphone type where both phones are stressed is very rare. This distribution of stress assignment is consistent for both languages, although the difference between the diphones with stress type 2 and 3 is larger in German than in English. The fact that some stress types in diphones occur more rarely than other might be used to reduce the database size to certain extent by covering for example only more frequently occurring diphone stress types.

Considering stress as context results in a total of 3,675 English and 3,981 German context dependant diphones which is a total of 7,656 diphones with stress. The total of English and German context independent diphones was 3,482. Thus already adding only stress for two languages increases the number of units in the database by 54.5%. If word accent for Bosnian was determined, the number of diphones to cover in the polyglot database would additionally increase. The increase would be higher than for English and German because instead of differentiating diphones based on stress, four word accents would had to be considered.

2.4.2 Syllable Boundary

Similar to stress, syllable boundary also has effect on the acoustic properties of diphones. The most significant acoustic change on the syllable boundary is drop in pitch, but this also can be followed by change in duration and amplitude as well (Saikachi 2003). Thus syllable boundary should be considered a possible context. Following (Saikachi 2003) eight possible syllable boundary contexts are defined, depending on the position of the diphone relative to the syllable boundary. Eight possible context variation means that the number of diphones would be multiplied by eight, if each possible diphone occurred in each context. This, however, is not the case. As already noted for stress the number of really occurring diphones relative to the syllable boundary is less than number of theoretically possible combinations. Frequency distributions of the eight syllable boundary positions relative to a diphone are given in table 2.5. Frequencies of single constellations show consistency among languages. Again, diphones occur in some syllable boundary contexts more frequently than in others. The most frequent context is having the diphone exactly at the syllable boundary, i.e. where

there is a syllable boundary between two phones. The context where, in addition to the syllable boundary between two phones, there is a syllable boundary to the left and to the right of the diphone occurs rarely.

Context	*Example*	*Type count*		
		Bosnian	*English*	*German*
1	A B	337	579	635
2	_A B	603	723	864
3	A B_	319	632	706
4	A_B	600	1260	1552
5	_A B_	362	559	801
6	A_B_	289	427	366
7	_A_B	747	492	446
8	_A_B_	286	116	55
Total		3,536	4,788	5,425

Table 2.5: Frequencies of different syllable boundary contexts for diphones._ marks the position of the syllable boundary relative to the model diphone AB.

The overall number of diphones when syllable context is added increases to 3,536 in Bosnian, 4,788 in English and 5,425 in German. In total, the number of syllable boundary dependent diphones is 13,749 which three times more than the number of context independent diphones in all three languages.

2.4.3 Position in the Intonational Phrase

One possible effect of the position within the intonational phrase on acoustic properties of diphones is phrase final lengthening. Syllabic segments (vowels and syllabic consonants) in the phrase final syllable have longer duration compared to the duration of same segments not in the phrase final position (Klatt 1975). These durational differences are perceptually relevant (Lehiste et al. 1976). Another possible type of alternation in the phrase final position is the change of the F0 contour when a declarative utterance is distinguished from a interrogative one, for example.

This implies that diphones in the phrase final position differ from the ”same” diphones (i.e. diphones involving same phones) in other positions in the phrase. This means

that both examples of diphones in the phrase final position and diphones not in this position should be stored in the database. Thus, phrase boundary context has two values, phrase final and phrase non-final. Table 2.6 gives frequencies of each of the two contexts for all three languages. The addition of phrase position with two values increases the number of diphones to 4,643 in Bosnian, 3,214 in English and 3,086 in German. This is a total increase of 6,554 diphones relative to the number of context independent diphones.

Context	*Example*	*Type count*		
		Bosnian	*English*	*German*
1	A B	3,702	2,515	2,558
2	A B#	941	699	498
Total		4,643	3,214	3,086

Table 2.6: Frequencies of different phrase boundary contexts (phrase non-final (1) and phrase final (2)) for diphones.# marks the position of the phrase boundary relative to the model diphone AB.

2.4.4 Single Language Cross-Word Diphones

Similar to the syllable and phrase boundary the word boundary can affect the acoustic realization of diphones. Changes in F0 contour and lengthening can occur at the end of a word, and depending on conversational situation also within a word (Saikachi 2003). Thus, word boundary can be considered an additional context according to which diphones should be distinguished. As for syllables, eight different context variations can be identified according to position of the word boundary relative to the diphone. Table 2.7 shows frequencies of each word boundary context.

Adding word boundary context parameters results in a total of 5,840 diphones for Bosnian, 6,257 diphones for English and 4,851 diphones for German. This is a total of 16,948 diphones. Compared to the number of context independent diphones which is 4,389, this is an increase of 74.1%.

Context	*Example*	*Type count*		
		Bosnian	*English*	*German*
1	A B	1,320	3,524	2,828
2	$A B	949	577	575
3	A B$	756	685	672
4	A$B	1,597	1,048	703
5	$A B$	274	199	67
6	AB	380	154	5
7	AB	492	62	1
8	AB$	72	8	0
Total		5,840	6,257	4,851

Table 2.7: Frequencies of different word boundary contexts for diphones.$ marks the position of the word boundary relative to the model diphone AB.

2.4.5 Cross-word Diphones Between Languages

Distributions of multilingual cross-word diphones, i.e. diphones at the boundaries of words from different languages, cannot be extracted from corpora since the no corpus contains enough language switches. Presumably, even if such corpus was available, the distribution of the multilingual cross-word diphones would potentially be very specific to the corpus. This is also the case for single language diphones, as noted above for different genres. However, since it there is little regularity in choice of words in code-switching strategies of single speakers, it can be assumed that the multilingual cross-word diphones which occur in one corpus frequently will occur more rarely in another corpus. In this case the cross-word diphone distribution would not be representative for the unlimited domain. However, further investigations across multilingual corpora are needed to confirm or reject this hypothesis.

Since frequencies of multilingual diphones cannot be extracted from corpora, full coverage of multilingual cross-word diphones is required, except in cases where these can be shared with the native langauge. The theoretical number of all diphones at Bosnian-English word boundaries is 1,380. This is the number of all possible phone-phone concatenations, with phones from the two languages. The same theoretical number of Bosnian-German diphones is 1,470. Other combinations of the three languages have

Bosnian	*English*	*German*
dc	dx	a
	ng	E
	hv	o
	nx	N
	el	e
	em	6
	en	x
	axr	

Table 2.8: Phones not occurring in word final position in Bosnian and in word initial position in English and German

not been considered in the project. Thus the only the case is examined where there is one change from Bosnian to another language. This is for example the case where the foreign word is the last word in the utterance.

Due to phonotactic constraints not all phones will occur at the end of a Bosnian word, nor all German and English phones occur word initially. However, this is true for only few phones. A list of phones not occurring in the word initial position in English and German and a list of phones not occurring in the word final position in Bosnian is given in table 2.8 (cf. appendix A for phonetic transcriptions). The reduced number of multilingual cross-word diphones is thus 1,209 for Bosnian-English words and 1,333 for Bosnian-German words. This reduces the initial number of theoretically possible diphones by 10.8%.

As pointed out in (Olive et al. 1998) stop, affricate and nasal combinations have minimal co-articulation properties, so they could be shared across languages without substantial spectral distortion. This means that if cross-word diphones include these phones and they already exist in Bosnian, they potentially do not have to be included again as cross-word diphones. The number of diphones which possibly could be shared is 364 for Bosnian-English and 426 for Bosnian-German word combinations. This would further reduce the size of the total cross-word diphone inventory from 2,542 to 1,752. This reduction, however, was not considered in further analysis since it has not been shown experimentally that sharing these diphones is indeed possible for the three languages. Thus the final number of context independent multilingual cross-word diphones is 1,209 for Bosnian-English and 1,333 for Bosnian-German language pairs.

The next step is to explore context variations of these diphones.

The number and possible variations of contexts in which diphones at the word boundary can occur are restricted. In all context types there will be a word and syllable boundary between the phones from different languages. This reduces the number of different context variations for syllable context from 8 in single language case (cf. table 2.5) to 4 as shown in table 2.10. Cross-word diphones including phones from stressed syllables can be differentiated from the ones without any stress. Boundary of intonational phrase does not make sense as a possible context for multilingual cross-word diphones. The only position of a crossword diphone relative to the phrase boundary is the one where the phrase boundary is between the phones (A#B). Since phrase boundary includes a break, there will typically be silence phones between the words at the phrase boundary. In this case, however, a multilingual cross-word diphone would not exist, but rather two diphones X_SIL and SIL_Y would exist, SIL being the silence, X the word final phone of the first word and Y the word initial phone of the second word. Thus stress, syllable boundary and word boundary are taken as possible contexts for multilingual cross-word diphones.

The table 2.9 shows the frequency distribution of context variations when stress is added to the English and German phones. The number of diphones increases to 7,482 which is 5,730 diphones more than when no context is considered. This number of different diphone types would increase further if Bosnian word accents were added.

Stress type	*Stress context type count*	
	Bosnian - English	*Bosnian - German*
AB’	1,769	1,624
AB	3,335	754
Total	5,104	2,378

Table 2.9: Frequency counts of stress contexts for multilingual cross-word diphones

Adding syllable boundary as context results in the total number of multilingual, cross-word diphone types of 15,257. The distribution of single context variations is given in the table 2.10.

If word boundary is taken as a context the variations presented in table 2.11 are possible. The context type number 2 is the case when a one-phone word precedes another

Syllable boundary position		*Context type count*	
Nr.	Example	*Bosnian - English*	*Bosnian - German*
1	A_B	2,960	3,737
2	A_B_	2,331	2,516
3	_A_B	1,092	1,372
4	_A_B_	644	560
	Total	7,072	8,185

Table 2.10: Frequency counts of syllable boundary contexts for multilingual cross-word diphones

word. Conversely, in the context 3 a one-phone word follows another word. Context number 4 is a word boundary diphone of two words containing only one phone. These short words (e.g. English determiner *a*, Bosnian conjunction *i* etc.) are very frequent. The Bosnian conjunction *i* (*and*) is, for example, the most frequent word in Bosnian corpus, occurring 20,980 times and the English determiner *a* is fifth most frequent word in English corpus, occurring 25,114 times. However, only few word types contain only one phone. In German corpus, no such words were found. Thus the number of possible multilingual cross-word diphone types at the boundaries of one-phone words is restricted. The most frequent word boundary context is context number 1, i.e. the standard case, where the boundary separates two words containing two or more phones.

Word boundary position		*Context type count*	
Nr.	Example	*Bosnian - English*	*Bosnian - German*
1	A$B	2,223	2,451
2	AB	117	129
3	AB	57	0
4	AB$	3	0
	Total	2,400	2,580

Table 2.11: Frequency counts for word boundary contexts for multilingual cross-word diphones

The total number of multilingual cross-word diphones when all three contexts, stress

for English and German, syllable boundary and word boundary, are added is 27,719. Compared to the total of context independent cross-word diphones, which is 1,752, this is an increase of 93.7%. The consequences of these single language and multilingual diphone distributions for design of a polyglot database are presented in the next section.

2.5 Construction of a Polyglot Database

The statistics and diphone distributions presented in previous sections were used to examine whether a database covering diphones from all three languages and additionally multilingual cross-word diphones can be constructed.

When single language diphone counts for all contexts (stress, syllable boundary, phrase boundary and word boundary) are added, the resulting total number of single language context-dependent diphones for all three languages is 45,122. Although this is already a large number, a more precise determination of how many diphones can be covered in the polyglot database is needed for discussion of whether a polyglot database with good coverage is feasible or not. As mentioned above, the size of the database is primarily constrained by the human capacities available for the recording of the database. This was taken as criterion for estimating how many diphones can be recorded for the polyglot database. This means that in order to estimate an acceptable number of diphones, it was necessary to know how long it would take to record the prompts covering these diphones. Clearly, this requires selection of prompts and their recording. Thus, the next step in analysis of design possibilities for the Bosnian-English-German polyglot database was to select prompts from the corpora and measure the time for recording them. A part of the prompts was used for building model voices for perception tests described in chapter 4. The selected prompts could also be used for building a polyglot voice at later stage. Building a full polyglot voice, however, was not the primary goal of this project, but it was envisaged for future work. At this stage the prompts are used to estimate which size of the polyglot database is acceptable.

In a polyglot database good diphone coverage has to be attained both for single language diphones and for multilingual cross-word diphones. In order to get this coverage, a set of sentences from single language corpora is automatically selected and recorded. The goal of the selection is to choose sentences which are representative for the whole corpus in terms of diphone coverage. Text selection can be done by a

commonly used greedy-algorithm proposed by (van Santen & Buchsbaum 1997). The algorithm weights sentences in accordance with their diphone coverage and thus selects an approximately optimal subset of a corpus which provides intended coverage. These sentences (prompts) are then recorded, and the database is constructed from recorded units.

Whole sentences, however, appear to be unsuitable for covering multilingual cross-word diphones, since no corpus has enough switches between languages within a sentence to provide sufficient number of multilingual cross-word diphones. A possible alternative is to greedily select sentences from the basic language, i.e. Bosnian with good coverage and then add words from foreign languages. The foreign words for the sentences would also be selected by the greedy algorithm, so that they cover required foreign language diphones. Foreign language words are positioned in the basic language sentences, so that the multilingual cross-word diphones are covered. This method was tried out for few sentences. The resulting nonsense sentences containing one or more code-switches were difficult to read. Apart from this, it was difficult to find a proper sentence intonation, when reading the sentences. Strange sentence intonation was introduced instead, which affected the words' acoustic and prosodic properties. In addition, inserting foreign words in a basic language sentence worked for few examples, but searching for the right place to insert the foreign word to get cross-word diphone coverage would be demanding if whole corpora had to be processed. Thus, a simpler alternative was chosen instead. It seemed more reasonable to have word pairs as prompts rather than whole sentences, when building a polyglot database. Words from single languages should be chosen to provide good coverage of single language diphones and the multilingual cross-word diphones can be covered at word boundaries.

Greedy algorithm was implemented to work on word types of single languages, rather than on sentences. A list of context-independent diphones to cover was defined. Uniform coverage of at least one occurrence of a context independent diphones was targeted. Thus, each word was assigned one score point for each diphone from the list which was covered in the word. After the coverage has been achieved, the diphone was deleted from the list of uncovered diphones, and the word was put on the list of selected words, sorted according to scores. The selection procedure ended when all diphones have been covered. The length of word list for Bosnian was 575 words, for English 920 and for German 1,152 words. The recording of these words would provide uniform coverage of context independent single language diphones. The next step was

to cover multilingual cross-word diphones. For this, Bosnian-English and Bosnian-German word pairs had to be built. Naturally, for building word pairs, first already selected words were chosen. The total possible number of multilingual diphones to be covered was 575 which is the number of selected Bosnian words. Bosnian words were first combined with English words. The combination resulted in covering only 87 cross-word diphones if each word is used only once. 488 words remained from the Bosnian list, since they wouldn't cover any new Bosnian-English diphones. These words were combined with German words. The coverage of uncovered cross-word diphones was low again (65 word pairs), since each word was used only once and cross-word diphones require several examples of same phones at the boundary. A total of 1,209 Bosnian-English and 1,333 Bosnian-German cross-word diphones had to be covered, so 2,390 multilingual cross-word diphones remained to be covered. Word selection procedure was run again. This time the aim was to provide Bosnian words ending in phones which are part of uncovered multilingual cross-word diphones. In analogy to this, English and German words starting with phones from these uncovered diphones have been selected. Arbitrary words from the three languages were selected. The first word which fulfilled the criterion of having the right phone in the end (for Bosnian words) or at the beginning (for English or German words) was chosen for each language and the word pairs were built. A better solution in this second run would have been to try to choose words with ”system”, so that more frequent diphones from the three languages are covered more than once for example. In this way, all 2,542 multilingual cross-word diphones have been covered. However, there were words left in all three languages which provided coverage for single language diphones but were not used in building multilingual cross-word diphones since their boundaries did not contain phones from diphones not covered. These were combined in arbitrary way and included in the database. There was a total of 1,039 such word pairs. Additional 21 word pairs were recorded to cover diphones needed in the experiment which could not be found in the database already. Exact selection criteria for these word pairs is described later in chapter 4, section 4.2.1. Thus the total of word pairs for recording was 3,602.

As previous analyses of context dependent diphone distributions show, the number of diphones to cover increases rapidly if any context parameter is added. Consequently, adding all contexts, stress, syllable boundary, word boundary and phrase boundary would require substantially more than 3,602 word pairs if all single language diphones

are to be covered. As an alternative to the uniform coverage of diphones, frequency weighted coverage as suggested in (François & Boëffard 2001, Saikachi 2003) could be applied to reduce the number of context dependent cross-word diphones and thus also reduce the recording time. These methods include removing diphones with frequency lower than 10, and could also include additional weighting of the words according to the frequency of the diphones covered in these words. As already mentioned, the main problem of frequency weighted coverage is LNRE property of languages because the probability that a diphone will not be found in the database at synthesis time increases.

The words pairs selected to cover context independent diphones were than recorded. As indicated above, acceptability of a database size was defined in terms of time needed for it's recording. For recording a set of 1,000 multilingual word pairs, approximately 1 hour was needed. Permanent switching between languages was difficult, especially for less common words, so breaks and false starts were made frequently during recording. What total time should be set for recording is an individual decision. However, the general guideline is that recordings should ideally be done on the same day to minimize the uncontrollable variations in voice quality. At the same time the recording procedure is very tiring, so it cannot be done on one day without losses in voice quality. For recording multilingual word pairs, maximal manageable recording time per day was 2 hours excluding breaks. It might be that a professional speaker, unlike myself, would be able to record longer, keeping the voice quality more constant. Given these initial observations a total recording time should not exceed 5 hours which means that a total of 5,000 cross-language word pairs was acceptable for the database. Word pairs covering context independent diphones, both single language and multilingual cross-word diphones, could be recorded. For context-dependent diphones, however, the number of diphones to cover is at least three times the number of context independent diphones for each context separately. This already would require unreasonably long recording time, so adding all context at once is clearly not feasible. However, context-dependent diphones are needed for unit selection database, so the best diphone for a given context can be chosen.

The multilingual cross-word diphones add to the problem in the case of context-dependent diphones. Frequency based selection for coverage of multilingual cross-word diphones is impossible since a corpus source for deriving the distribution is not available as mentioned in section2.4.5. The conclusion there was that all multilingual cross-word diphones must be included in a polyglot database. When reductions suggested in section

2.4.5 are made, there is still a possible total of 27,719 context dependent multilingual cross-word diphones to be included in the database in addition to single language diphones. In terms of recording time, as calculated above, covering this number of diphones in the polyglot database is not achievable. It should also be noted again that this is only the number of cross-word diphones where the first word is Bosnian and the second word English or German. A real world unlimited domain polyglot synthesizer has to be able to handle arbitrary language combinations which means even more cross-word diphones to cover. To reduce this number, not all contexts could be selected at the same time. What influence adding different context would have on the output is an interesting question which remains to be investigated.

Thus, even though the frequency based coverage of single language context-dependent diphones might be possible by using frequency based methods described in literature, adding multilingual cross-word diphones exceeds the database size beyond acceptable limits. An attempt to reduce single language coverage further and add some multilingual cross-word diphones would be a compromise leading to a database with insufficient coverage for both single language and multilingual cross-word diphones. A better solution to the polyglot database design would be to provide as good single language diphone coverage as possible and to deal with multilingual cross-word diphones by an alternative method at synthesis time, instead of including them in the database.

2.6 Summary

In this section possibilities of building a polyglot database with good unit coverage for all three languages Bosnian, English and German were explored. It was shown that diphones are the best units in terms of compromise between unit coverage and database size. The number of context independent diphones needed for coverage of the polyglot database is 4,389. For unit selection however, context-dependent diphones are needed. It was shown that it is impossible to cover all context-dependent diphones for unlimited domain databases even in a single language case. The number of context-dependent diphones for contexts stress (for English and German), syllable boundary, phrase boundary and word boundary is 45,122 which is prohibitive in terms of recording time. An approximation to good coverage might be made if only high frequency diphones are covered as previously suggested in the literature. However, although frequency weighted coverage of context-dependent diphones is reported to be superior

to the coverage of context independent diphones it is not a good approximation because of the LNRE properties of speech. For polyglot databases there is an additional problem of multilingual cross-word diphones. 2,542 context independent multilingual cross-word diphones can potentially occur in the three languages. Even if no context is considered for these cross-word diphones, adding them to the context independent single language diphones for the polyglot database increases the number of units to cover to 6,931. In the context dependent case the problem of database size multiplies. Given these distribution facts, it is worth exploring alternative possibilities of dealing with multilingual cross-word diphones without including them in the polyglot database. Several alternative approaches to handling these diphones are described in the next chapter.

Chapter 3

Approaches to covering multilingual cross-word diphones

As shown in previous chapter a single polyglot database with good coverage of single-language context-dependent diphones which additionally provides coverage of multilingual cross-word diphones is not feasible. The aim of this chapter is to describe alternative possibilities of dealing with multilingual cross-word diphones in speech synthesis.

All approaches to handling multilingual cross-word diphones can be divided into three groups according to the extent to which the these diphones are covered in a basic language database. The first solution is already described in previous chapter. It includes full coverage of one example of each multilingual cross-word diphone, i.e. context-independent coverage of multilingual diphones, and was shown not to be feasible in general. Second, the coverage can be partial, so that only those foreign phones, substantially different from the basic language ones are covered. Finally, there are different methods for synthesizing speech from databases with good coverage for single languages, but without any inclusion of multilingual cross-language diphones at word boundaries. Both full coverage of one example of a diphone and alternative methods of database design were applied and the resulting quality of synthesized speech was tested. The testing procedure and results are described in the next chapter. This chapter will point out some problems with all approaches to design of a polyglot database. The problems are illustrated on the examples from the testing material described in section 4.2.1.

3.1 Full coverage

One possibility to handle cross-word diphones at the boundary of two words from different languages is the attempt to cover one example of each cross-word diphone for all language pairs. This means that phonetic or prosodic context are not taken into consideration, comparable to building a database for diphone synthesis.

Having a diphone in the database results in good synthesis quality. Naturally, the best quality of output speech is achieved if the units are recorded together. The quality of concatenation of units not recorded together depends primarily on the type of the diphone and also on the phonetic environment in which the diphone was recorded. Olive et al. (1998) give a list of consonant pairs which exhibit minimal coarticulation on each other. Stop-stop combination is an example of such consonant pair. It can be expected that concatenation of these phone pairs with minimal coarticulation has better quality than synthesis of diphones including vowel combinations for example since vowels are known to have strong co-articulation effects on their environment.

The concatenation of stops is illustrated in figure 3.1. The figure shows the spectrograms of Bosnian-German word pairs ”izlog Partner” (*shop window partner*) and ”prilog Partner” (*contribution partner*). The first word pair is recorded natural speech and the second is synthesized speech. The word ”prilog” was recorded in the word pair ”prilog Parade” (*contribution parade*). In the spectrogram the word boundary is marked with the ellipse. As the marked part of the spectrogram indicates, there is very little difference in the shape of the spectrum at the word boundary as it could be expected for the stop-stop combination. Thus the synthesis output sounds very close to the recorded speech.

The example illustrated in figure 3.1 also shows how context contributes to the synthesis quality. The phonetic and prosodic context to the left and to the right of the cross-word diphone /g p/ in this example is similar. The diphone /o g/ for synthesis of ”prilog” is taken from the same word as recorded. The diphone /o g/ from the word ”prilog” is almost the same as that from the word ”izlog” since the contexts for the diphone /o g/ in both words are similar. Both diphones are preceded by /l/ in the word-final position and in the unstressed syllable. The cross-word diphone /g p/ comes from the word pair ”izlog Partner”. Thus there is a join in concatenation of /g p/ to /o g/. However, it is not audible which is also due to the similarity of environments. On the right hand side the cross-word diphone, the diphones for the word ”Partner” are used

as recorded in the word pair ”izlog Partner”. Thus the recorded context and the context for synthesis remain the same for the word ”Partner”, so overall synthesis quality of the word pair ”prilog Partner” is very good.

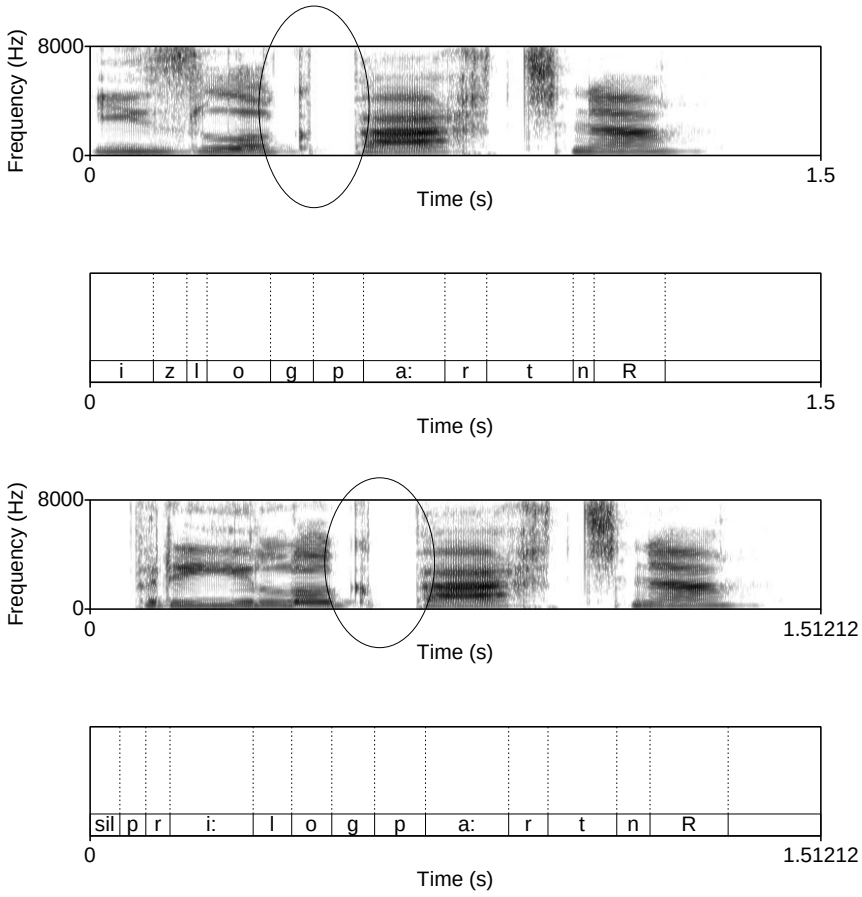

Figure 3.1: Spectrograms of the recorded word pair *”izlog Partner” (shop-window partner)* (top) and synthesized example *”prilog Partner”* (bottom) show no substantial differences in the spectral shape at the word boundary

Figure 3.2 on the contrary illustrates to some extent the situation which is problematic for this method. The figure shows the spectrogram of the Bosnian-English word pair ”šarafić that” (*small screw that*). The cross-word diphone /tc dh/ is recorded in the word pair ”čekić themselves” (hammer themselves). However, the diphone /f i/ is taken from another recording in the database, so the concatenation of /f i/ and /i tc/ results in spectral distortion and audible join to the left of the cross-word diphone. This also illustrates the point that concatenation of vowels is more problematic than that of stops even in similar environments. Thus including only one example of a multilingual cross-word diphone containing a vowel can result in bad quality of concatenation to

the preceding or following diphone. On the right hand side of the cross-word diphone /tc dh/ there is no join since both diphones /tc dh/ and /dh eh/ are taken from the same recording.

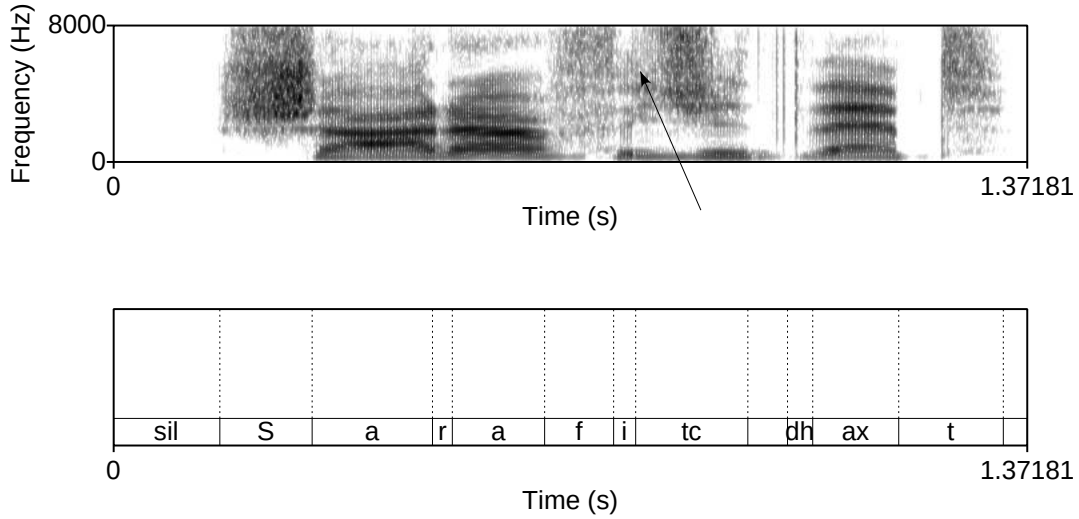

Figure 3.2: In the spectrograms of the recorded word pair *"šarafić that"* there is an audible join between diphones /f i/ and /i tc/

Recording cross-word diphones in only one context causes discontinuities if the diphone is used in a context other than the recorded one. Depending on the type of phones involved, these discontinuities are more or less audible as the examples illustrate. Using a diphone recorded in only one fixed context for synthesis of arbitrary contexts means backing-off to diphone synthesis whenever a foreign word is encountered in the input. Strategies involving resorting to diphone synthesis where unit selection does not work have already been tested (Stöber et al. 1999) and reported to result in poor overall quality, probably due to striking variation in quality within the same utterance. A unit selection database should contain several examples of a unit in order to choose the one with lowest join costs. If there is only one example of a cross-word diphone in the database, the join costs to the units to the left and right from the cross-word diphone will potentially be high for word combinations other than the ones recorded together, thus resulting in audible joins and lower quality.

At the same time, as already mentioned in the previous chapter, the number of context independent multilingual cross-word in is large. For the Bosnian basic voice which includes German and English words the total number of context independent cross-word diphone types is 2,542 for Bosnian-English and Bosnian-German combinations. This

number of cross-word diphones means approximately two and half hours of recording multi-language word pairs according to calculations in chapter 2. If other combinations within three languages are added, i.e. if also English-Bosnian and English-German and German-English diphones are considered, the database easily increases beyond the limits of feasibility.

Increase in number of multilingual cross-word diphones to be covered goes along with more complicated procedures of text selection. For calculation purposes in previous chapter, arbitrary words pairs were built if they cover a cross-word diphone. However, a systematic selection procedure is required for attaining a compromise between database size and good diphone coverage. For more than three languages all these problems multiply, so that covering all cross-word diphones across languages becomes impossible.

3.2 Databases with single language coverage

Almost all speech synthesizers are multilingual, i.e. have several single language databases for different monolingual voices. These can be used for polyglot synthesis. Alternatively, a single polyglot database could be constructed to include only good coverage of single language diphones. Here, the second case will be considered. When the textual input requires synthesis of a native-foreign cross-word diphone where the diphone is not in the database, there are three possibilities to synthesize cross-word diphones from one or more single language databases. These are described in the following sections.

3.2.1 Full nativization

The first possibility is the attempt to accommodate a foreign phone in a cross-word diphone to the closest native phone as suggested in (Badino et al. 2004). The main problem with this way of handling foreign diphones occurs in the cases where a foreign language phone does not exist in the basic language phone inventory. In our system this is the case when German or English phone does not exist in Bosnian. Replacements by the closest Bosnian phone can render unintelligible or inappropriate synthetic speech.

This problem can be illustrated on the Bosnian-German word pair ”pomoć Pflüge”

(*help ploughs*) from the test set. The affricate /pf/ is not a part of the Bosnian vowel inventory. On the other hand the affricate /tc/ at the end of the Bosnian word also does not exist in German. Thus the only way to cover the diphone /tc pf/ is to include a multilingual language pair in the database. Since our database should only have single language coverage no instance of the diphone /tc pf/ can be found in the database. In the case of nativization of phones not in the database, this means that a closest match for the affricate /pf/ in Bosnian is searched for. Badino et al. (2004) define the closest match as the result of an automatic search for the most similar phones based on weighted perceptual similarity. The definition of phone similarity here is based on informal productional tests, i.e. on the question, how a sound not in the phone inventory of Bosnian would most likely be realized by a Bosnian native speaker who fails to pronounce the foreign phone as a native speaker of foreign language. For German affricate /pf/, the closest match in Bosnian would most likely be the fricative /f/. The diphone created after the matching is /tc f/. Synthesized utterance renders ”pomoć Flüge”. /pf/ and /f/ contrast word initially in German (e.g. in the minimal pair ”Plüge” (*ploughs*) vs. ”Fl”uge” (*flights*)) so the synthesis is not appropriate. Thus the main problem of nativized pronunciation is that it can alter the meaning of a word and make it semantically or syntactically inappropriate for a given sentence context.

3.2.2 Phone concatenation

Second way to handle foreign cross-word diphones not in a database with single language coverage is to resort to phone concatenation. That means for example that if a cross-word diphone can not be found, the word final Bosnian phone and word initial German/English phone are concatenated.

In analogy to what has been reported in (Stöber et al. 1999) for resorting from unit selection to diphone synthesis, spectral mismatches and joins in the middle of the concatenated phones could be expected if the synthesizer backs off from diphones to phones. In figure 3.3 this problem can be seen on the example of the synthesized word pair ”tutanj Viertel” (*roar(n.) quarter*).

Apart from this quality problem phone concatenation can have another problem affecting intelligibility. The algorithm which concatenates phones used in this practical is activated if a cross-word diphone cannot be found. It extends the last phone of the first word (Bosnian word) to the right and the first phone of the second word (English or

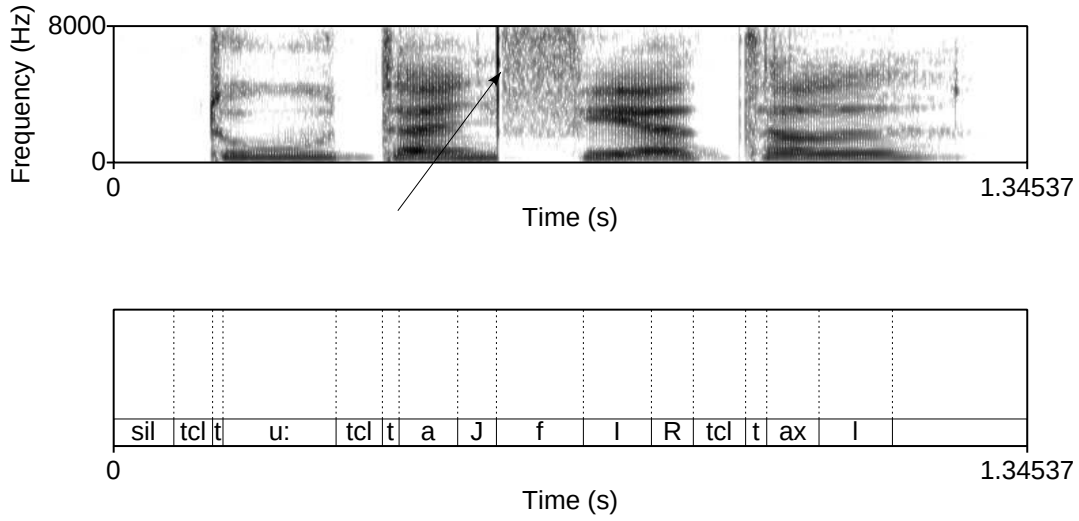

Figure 3.3: The spectrogram of the word pair *"tutanj Viertel"* shows a join between two words when phone concatenation is used

German word) to the left and thus concatenates the two words. This requires precise labelling on the phone level. Accurate detection of phone boundaries is a problem even for humans, as evaluations in (Makashay et al. 2000) suggest. Unsteadiness of phone boundaries and co-articulation effects make it difficult to judge where one phone ends and the next one begins. Automatic methods are expected to be at best as good as humans on this task, so even more inaccurate phone labelling is expected. Forced alignment used for labelling here is reported to be ”consistent and reasonably accurate” (Clark et al. 2004). It does however, introduce labelling errors (cf. section 4.2.2.1 in chapter 4 on labelling procedure).

3.2.3 Inserting a pause

A simple solution to the problem of covering cross-word diphones without extending the database is to separate the words from different languages by a short pause. Spectral discontinuities may thus be hidden in the silence and the overall speech quality might appear better. At the same time, the listeners might expect a short pause between the words, so that this generally does not sound unnatural.

How natural the word pair with inserted pause sounds, depends primarily on the length of the pause. Waveform and spectrogram of the word pair ”konac appoint” (thread

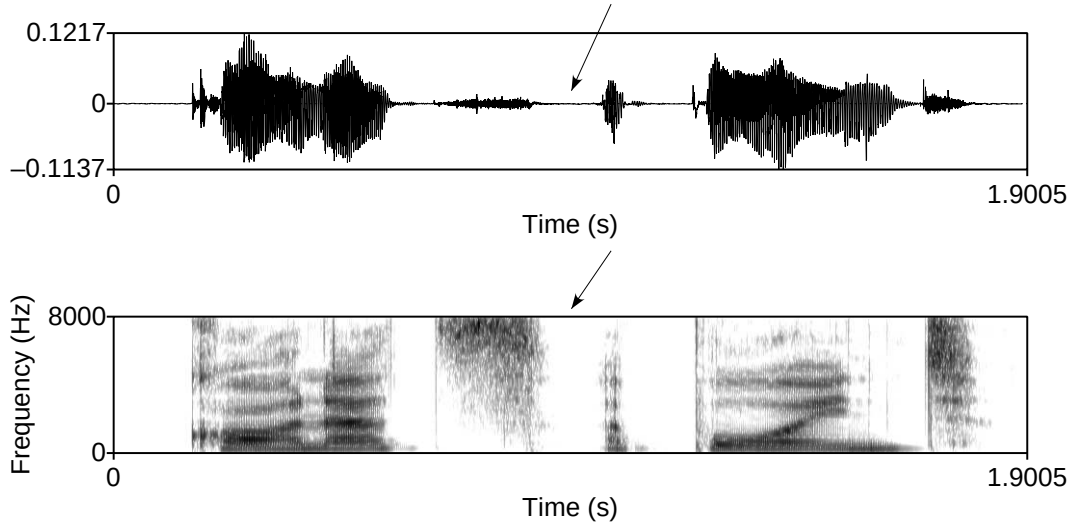

Figure 3.4: Waveform and spectrogram of the word pair *"konac appoint"* shows long pause between the words when an extra pause is inserted

appoint) in Figure 3.4 show a pause sounding almost unnaturally long. This problem, however, can easily be fixed by setting a maximum length of a pause between two words and cutting off too long periods of silence to fit this maximum.

As in phone concatenation, a problem can appear if there are errors in automatic labelling of the pause (cf. section 4.2.2.1, chapter 4). For pause the problem is that additional elements can be introduced along with the pause. Figure 3.5 shows three spectrograms of the word pair ”vrtlog Parfüm”. The bottom spectrogram shows the word pair when pause is inserted. Compared to the recording of the same word pair (top) and it synthesis without a pause (middle) the word pair with the pause has some creaky voice content after the first word which affects the overall quality. This shows that labelling errors are general problem in annotation of large databases. They can impair synthesis quality of any method for handling multilingual cross-word diphones.

Some sound type combinations are more suitable for inserting a pause between them than the others. A pause between stops, or affricates is might not be as perceivable as the pause between more continuous sounds. At word boundaries where more continuous sounds like vowels, glides or fricatives come together the pause can sound like an unnatural break. In the sentence context pause like this might interrupt the fluency of the sentence. For word pairs, the problem with the pause in the vowel context is that the join between the vowel and pause is more audible than in when a consonant is

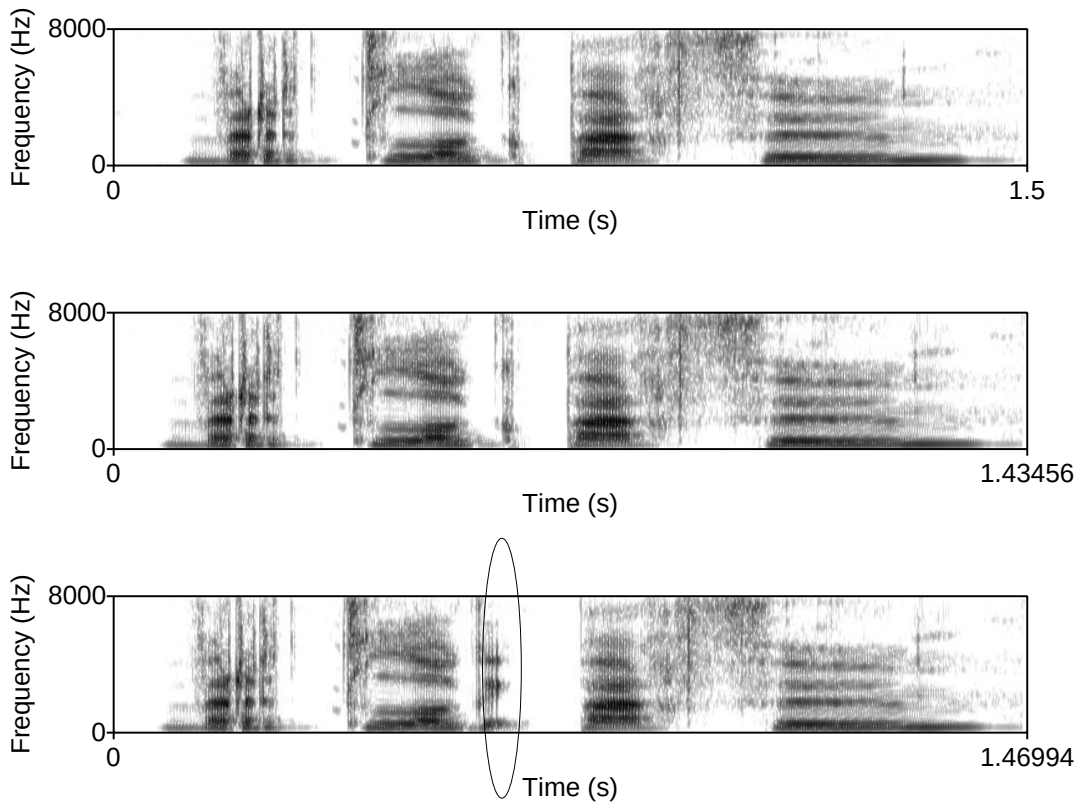

Figure 3.5: Spectrograms of the word pair *"vrtlog Parfüm"*. Recorded word pair (top), the synthesis without pause insertion (middle) and the pause insertion (bottom) and . The last spectrogram shows some creaky voice content between the words.

concatenated to the silence. The spectrogram of the word pair "kraju append" (*(at the) end append*) in figure 3.6 illustrates this problem.

The unstressed schwa vowel at the beginning of the English word exists as it is clearly visible in the spectrogram. Nevertheless, to several listeners in the informal testing it sounds like extended join, so that several listeners understood "kraju pend", rather than "kraju append".

3.3 Partial coverage

Instead of trying to cover all possible cross-word diphones in all languages or of having only a single language database a combination of the two approaches can be made. In this case the goal is to cover only those diphones which include phones not in the

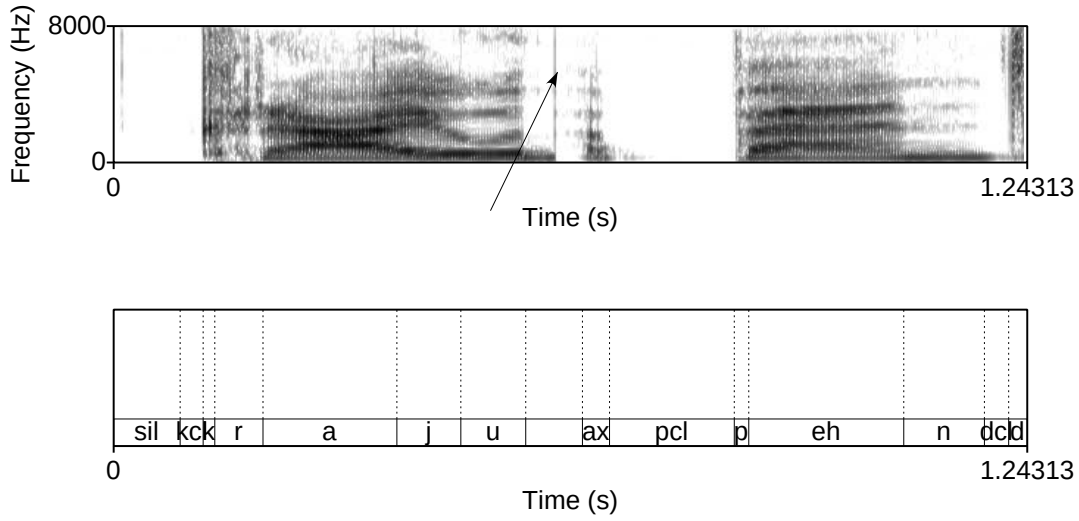

Figure 3.6: Spectrogram of the word pair *"kraju append"* shows join between in the diphone /u sil/. The join is audible and the following schwa vowel is not always audible.

phone set of the basic language and approximate all other diphones with basic language units. For the three languages database this might be a good approach. Single language coverage could be high as in the approaches described in previous section. Additionally, multiple examples of cross-word diphones across languages which are problematic for the single language approach could be covered.

This approach also combines the problems of the two other approaches. It has the same problem as full coverage approach in cases where the foreign diphone is used if there is not a good example in the database. However, the output speech should have less joins within language words because more single language diphones are employed in synthesis and these concatenate smoother. Usage of similar units for different languages makes the voice sound more nativized to a particular language. However, defining the minimal set of phones not in the language is not straightforward. Thus deciding which units to cover in the database and which to handle by an alternative method requires a lot of planning and knowledge of the phonetics of the languages involved.

3.4 Summary

In this chapter several approaches to handling inter-language cross-word diphones were described. In chapter 2 it has been shown that full coverage is generally an

unfeasible option. At the same time single language databases can be built to cover context-dependent diphones to reasonable extent as argued in the literature and also shown in the previous chapter. This being the case it is worth testing whether producing multilingual cross-word diphones on synthesis time by concatenating units from single language databases renders same or better quality synthesis as covering all multilingual cross-word diphones once in the database. The next chapter reports on evaluation of the approaches described in this chapter.

Chapter 4

Evaluation

The aim of this chapter is to evaluate different methods of database design for handling cross-word diphones in words from different languages in a polyglot unit selection synthesizer. One of the major goals is to show that full coverage of cross language diphones in a unit selection database is neither feasible and desirable nor necessary. The main hypothesis is that cross-word diphones in words from different languages synthesized from single language databases either by concatenating phones or by inserting a pause between two words from different languages sound at least as intelligible and natural as the speech created from databases with full cross-language diphone coverage. This means that in order to build a polyglot voice the database does not have to be extended by foreign language sounds but can be effectively created from databases with single language coverage without loss in spectral smoothness.

4.1 Goals

The description of methods of handling multilingual cross-word diphones described in chapter 3 suggests that each method has potential drawbacks and none would render good synthetic speech in all cases. The methods, where no database extensions are needed are better in that less resources are required for the database construction. The aim of the following experiments is to show that these methods also render same or better quality speech than the methods including extension of the database which is an additional reason why these should be preferred in the polyglot speech synthesis.

There is no general agreement on standards in speech synthesis evaluation. How-

ever, most synthetic voices are evaluated for their intelligibility and naturalness using listening tests with human listeners. This evaluation strategy is also adopted here. Four methods for handling cross-word diphones in words from different languages are tested for intelligibility and naturalness of speech produced when these methods are employed. Full coverage of one example of each multilingual cross-word diphone served as baseline. Three methods involving units from single language databases: full nativization, resorting to phone concatenation and inserting a pause are compared to the baseline. The hypothesis is that either resorting to concatenation of phones or inserting the pause renders same or better quality synthesis than including one example of each multi-language cross-word diphone in the database. It is also investigated how full nativization of foreign units is accepted by the human listeners.

4.2 Methodology

The evaluation goals outlined in the previous section first required a definition of a set of utterances to be submitted to the subjects for listening. Then, model databases have been developed and four synthetic voices have been built. Each of the voices employed different method of handling multilingual cross-word diphones. The test utterances were synthesized and presented to the bilingual subjects for intelligibility and naturalness evaluation. Finally, the results were evaluated and analyzed. The following sections describe the evaluation process in detail.

4.2.1 Testing materials

The four methods for handling multilingual cross-word diphones were tested on Bosnian-English and Bosnian-German word connections. For evaluation, word pairs rather than larger phrases or sentences were used. The main reason for using word pairs is that the evaluation of the spectral quality at the cross-language word boundary is difficult in the longer units of speech because subjects might also evaluate spectral quality of other parts of the utterance not only of the target word boundary between two foreign words. Furthermore, the voices were built from the word pairs database. Word pairs mostly have list reading intonation which would sound unnatural in the utterances. Thus using word pairs instead of whole sentences also prevents introduction of additional influence of unsuitable prosody on subjects' judgements.

All words used in the word pairs were disyllabic, except two English monosyllabic words which were used since suitable disyllabic words could not be found. It was important to keep the word length constant. Otherwise it could be an additional confounding factor in intelligibility judgements, since shorter words are more difficult to understand than longer ones. Disyllabic words were chosen because it was more difficult to find enough monosyllabic word examples in the corpora so that the required phone categories described below are represented.

The target of the evaluation is only the quality of the cross-language word boundary. Thus it was important to avoid further joins between single diphones within the words building a word pair, since these might be the reason for a particular quality judgement, rather than word boundary diphones. This was achieved by including all the words needed for testing in the database. In the synthesis the units recorded together are chosen since they have the least concatenation cost. Thus the within word synthesis quality is kept close to the recorded speech, so that it can be assumed that only word boundary quality influences judgments.

The main decision in choice of the testing utterances is the choice of the cross-word diphones to test. Testing all cross-word diphones is impossible, and even if only combinations between phone groups (i.e. vowels, stops, fricatives, nasals etc.) are taken, this results in a vast number of combinations. In fact, having more than approximately 50 word pairs for both intelligibility and naturalness test did not seem to be recommendable. When assessing the naturalness of the speech subjects' judgements tend to become more similar after certain number of heard examples. The quality of different synthesized examples tends to be perceived as same. In the assessment of short word pairs it is particularly probable that the judgements will converge if there are many word pairs to assess. Thus it seemed to be advisable to limit the number of the examples presented to the subjects. This practical constraint drastically limits the number of diphones on which different methods for synthesis of cross-word diphones can be tested.

Four Bosnian-German and Bosnian-English cross-word diphones were chosen for testing. For Bosnian-German the test cross-word diphones were: /g p/, /J f/, /o E/ and /tc pf/ and for Bosnian-English word combinations: /g p/, /J f/, /L th/, /ts ax/ (cf. appendix A for transcriptions and IPA symbols). Why these diphones and not some others? The attempt was to cover as many sound classes as possible, i.e. to have an example of a stop, fricative, affricate, vowel etc. A further criterion was to represent certain problem

cases. As outlined in chapter 3 each method can render bad synthesis on certain sound groups. Stop-stop combinations, like /g p/ for example are considered less problematic for all methods, than vowel-vowel combination /o E/. In the first diphone both phones are in all languages which is important for nativization approach, and the coarticulation between them is minimal which affects phone concatenation. Also coarticulation effects to the left and right of the cross-word diphone are expected to be less strong than coarticulation effects of vowels. If this holds, than covering one example of a stop-stop diphone and using it in another context would render better synthesis than doing the same for vowel-vowel diphone. Similar to /g p/ the diphone /J f/ is considered unproblematic for all methods. The diphones /tc pf/, /L th/, /o E/ and /ts ax/ on the contrary are expected to be problematic. They present a problem for nativization approach since they include phones which differ across languages. They also include phone classes where more coarticulation is expected between the phones and at the phone boundaries. This makes them potentially problematic for phone concatenation and inclusion in the database in the context other than the recorded one. The synthesis of both problematic and unproblematic cases should be represented in the testing material in order to avoid having better synthesis output only because the diphone is unproblematic for synthesis.

Each diphone was synthesized by each cross-word diphone handling method, so that the effectiveness of different methods could be compared. The problem here was that listeners probably will remember a word pair if they hear it more than once, so intelligibility judgement will be confounded by guessing the word-pair from what is in the listener's memory. Naturalness assessment can also be influenced by this in that the same judgement is given to the a word pair synthesized by different methods, because subjects can remember what mark they already have given to this particular word-pair. Thus it was important to prevent that subjects hear same word pair more than once. At the same time same, diphones had to be synthesized with each of the four different test methods. The solution was to embed same cross-word diphones in different word pairs and each word pair was synthesized by a different method. In this way, the quality of the synthesis for each of the four methods in each word pair could be assessed. However, this solution to the priming problem was not optimal. The problem with this is that the realization of a cross-word diphone, which is supposed to be the same in all different word pairs, is in fact not exactly the same due to co-articulation effects of the neighboring sounds. Thus, another confounding factor is potentially in-

troduced by using different word pairs for testing same cross-word diphones. This problem is minimized by trying to keep the environment of each diphone as similar as possible. All word pairs having a particular diphone at the boundary should have same phones surrounding the diphone and same stress. Finding four disyllabic words with same phones around the diphone and same stress was not always possible. In the cases where no suitable word pairs could be found the preference was given to the similarity of phonetic context over stress. A list of all tested word pairs is given in appendix B.

Word pairs were selected manually and automatically from the three single language corpora. First, all possible words with target diphones at the word end for Bosnian and at the beginning of the word for English and German were selected. Then only disyllabic words were filtered out using lexicons for English and German and syllabification rules for Bosnian. Finally, the lexicons were also used to check the stress placement and find the words with same stress to provide the same environment for the word final phones as discussed above. Words with same stress were easy to find for English but could not be found in many cases for German. Thus the constancy of the environment in which a phone occurs is not always given for German. Since there is no stress information for Bosnian, a simplifying assumption was made that all disyllabic words have stress on the first syllable. Than, Bosnian words for testing were selected according to phonetic environment. It appeared that the assumption about stress assignment was confirmed, at least for all selected test words.

If two words fulfilled all criteria, but one of them contained more frequent diphones than the other, the word with more frequent diphones was chosen. The reason behind this is that the less frequent the diphones in the word are, the more likely the subjects are to recognize the word based on the peculiarity of the sounds in it. An example of this are Bosnian words ”toranj” (tower) and ”žrvanj” (millstone). Both of them contain the final phone /J/ and would be suitable for testing the ”/J f/” cross-word diphone. However, the word ”žrvanj” contains more unusual phone combinations, so that the subjects might guess it based on their knowledge that not many words apart from this one sound like that. Generally, more common words were preferred to unusual and archaic words even if the latter suited better considering other criteria. It was important to ensure that subjects are familiar with all words they hear as far as this was possible due to limited number of phonetically suitable disyllabic words. In this way, wrong intelligibility judgements because the word is not known should be avoided.

The final criterion in word choice was to choose words with neutral semantics and

avoid for example all negatively coloured or embarrassing words which subjects might ”choose” not to understand because of some social constraints.

4.2.2 Building the voices

The test utterances described above were synthesized by three different synthetic voices. The voices were built using voice building tools for Festival's new unit selection engine, *multisyn* (Clark et al. 2004). These voice building tools are based upon Festvox voice building tools (Black & Lenzo 2003). Additionally, HTK speech recognition toolkit (Young et al. 2002) was used to do automatic labelling. The three voices are built from the model databases containing only utterances needed for the evaluation. They differ in specification of the way how to handle cross-word diphones.

The process of building a new synthetic voice is almost fully automatized. After recording, the prompts are automatically labelled on the phone level by forced alignment. Generally, the synthesizer front-end had to be customized if the prompts contain special symbols or more elaborate phrasing. Here however, this is not the case, since no abbreviations or symbols are used, and also whole word pairs can be seen as one intonational phrase. After labelling, voice building was done automatically by the tools provided in voice building tools for *multisyn*. The following sections describe steps in voice building.

4.2.2.1 Forced Alignment

Instead of labelling the utterances manually, methods used in speech recognition were employed to align recorded prompts with their phonetic transcription. Alignment can be viewed as a simplified recognition task, where the sequence to be recognized is known. Here, HTK speech recognition toolkit based on Hidden Markov Models (HMMs) (Young et al. 2002) was used to do the alignment. The labelling procedure used here is described in (Clark et al. 2004). It involves preliminary labelling using Festival text processing front end, where a sequence of segments needed for transcription of the written prompts is generated by lexical look up. Stop and affricate closures, short pauses between words and utterance initial and final silence are then added. The next step is to build monophone HMMs for each phone and train them using this initial transcription. The model parameters are re-estimated in four iterations, and the phones

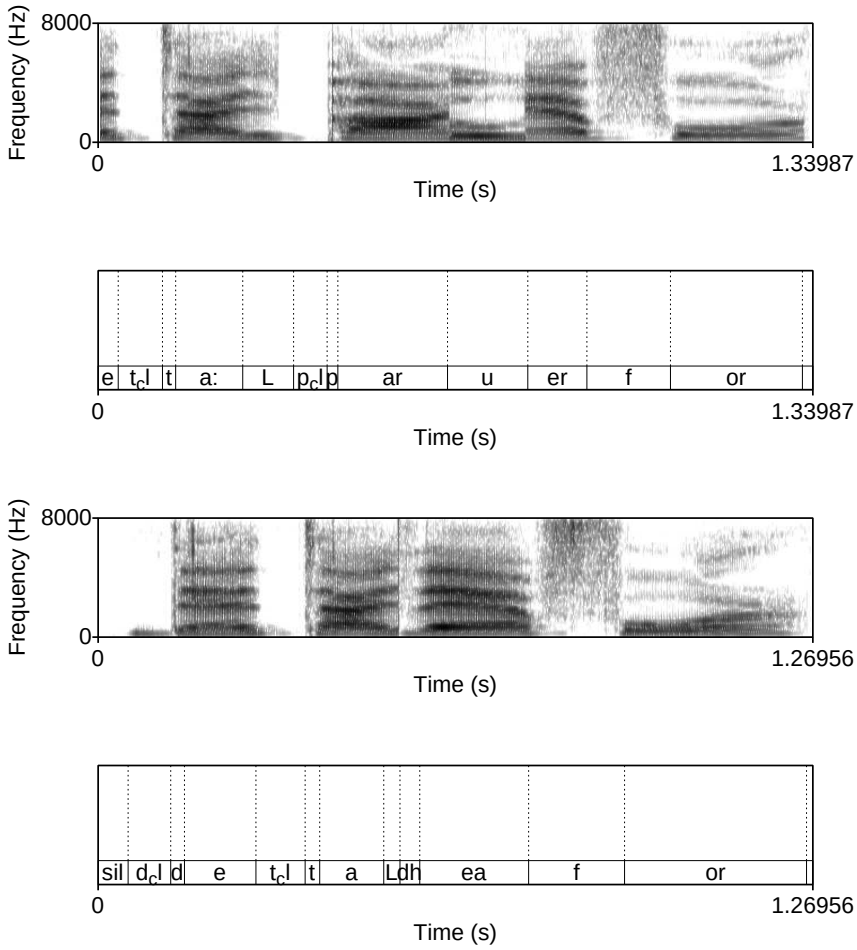

Figure 4.1: Spectrogram of the word pair *"detalj therefore"* when synthesized from automatically labelled units (upper) differs substantially from the same synthesized word pair after manual correction of phone labels

and transcriptions are then realigned. In the next step, the number of Gaussian mixtures in the models is increased from one to eight. Then, the final re-alignment is done which results in the labels for the speech database.

Errors in labelling substantially affect the intelligibility and quality of synthesis. The input text is synthesized with wrong units which are, due to their wrong labels, mistaken for units suitable for certain word. Figure 4.1 illustrates the problem. The upper spectrogram shows the word pair ”detalj therefore” (*detail therefore*) synthesized from automatically labelled units. In the spectrogram the initial stop /d/ is missing and so does the first part of the word ”therefore”. Instead, additional sounds are inserted between two words. The lower spectrogram shows the same word pair after manual

correction of labels. Both words in the word pair are now properly labelled and the synthesized speech is intelligible without missing or wrong units in it.

4.2.2.2 Extracting pitch marks

After labelling the prompts, pitch marks were extracted. Pitch marks are used for pitch synchronous linear prediction analysis. In *multisyn* linear prediction (LP) coefficients and residual are calculated from speech frames separated on pitch marks, rather than on fixed time intervals. If the vocal fold movements of the speaker are recorded with an electroglottograph (EGG), the pitch marks can be extracted from EGG signals during recording of the prompts. Since this was not the case here, the pitch marks had to be extracted from the waveform itself.

After extracting, pitch marks are first automatically corrected by moving each pitch mark to its nearest peak. The pitchmarks were than examined using XWaves and further corrections were made. The first correction applied to the pitchmarks was to change the minimum and maximum values to reflect the frequency range for female voice. This brought already significant improvements in pitch-marking. Next, the cut off frequencies for high and low pass filters were adjusted as suggested in (Clark & King 2003). This resulted in better pitch-marking, so no further adjustments were necessary.

4.2.2.3 Building utterance structure

The next step in building the voice was to use Festival to create utterance structures of the prompts. Utterance structure stores linguistic information about each utterance in the database. It consists of a set of items representing objects like word, syllable, segment which are organized by relations like Segment, SylStructure, etc. Linguistic information needed for unit selection synthesis includes information on segments (i.e. phonetic transcription) and phrasing. Prosodic knowledge, like duration and F0 contour, which have to be modelled in the diphone synthesis, are taken from the database of recorded speech.

Utterance is also the basic unit of synthesis in Festival. In the synthesis process, first, a target utterance structure is created. Than, candidate units are chosen from the database

to fit each target unit defined in the utterance structure. Finally, the best candidate units sequence is chosen if it minimizes target and join costs.

4.2.2.4 Duration and F0 contour

In the labelling, each diphone is labelled with its start and end time. If automatic labelling goes wrong, segments might be assigned too long duration. To avoid this, distribution of segment duration is computed, and outliers with much longer duration are marked in the utterance structures. These units are not used in the synthesis. In the synthesis, duration information is used in target costs, where diphones with more natural duration are favoured.

Next, F0 pitch track contour is generated. Again, minimal and maximal pitch values had to be changed to the values suitable for female frequency range.

4.2.2.5 MELCEP parametrization

Mel Frequency Cepstral Coefficients (MFCCs) are parametric representation of speech which represents waveform as a vector of numbers which are not correlated with each other. This is a useful property for statistical models of speech since it reduces the number of parameters needed for modelling. At the same time, MFCCs reflect the non-linear relationship between frequency and pitch as perceived by humans. MFCCs are created at the beginning of the voice building process because they are used for training HMMs for forced alignment. In this final step in voice building MFCCS are normalized to lie within the range [0,1], and these normalized MFCCs are used for calculation of the join costs (Clark et al. 2004).

4.2.3 Voices and Synthesis

Three different voices were built using the procedure outlined above. These are described in the following section along with synthesis methods used in each case. Which word pair was synthesized by which method is shown in the appendix B.

Full Coverage and Insertion of a Pause

The first voice *(multiling_full_pause_multisyn)* was built from the database given in appendix C.1. The appendix shows the recorded prompts selected for this voice. It was used to synthesize utterances with method FULL (meaning: full coverage of one example of a cross-word diphone in the database and synthesis of the same diphone in another context) and method ”PAUSE” which includes inserting a pause between two words. In both cases all diphones needed for synthesis were planned to be in the database, so no action was needed for handling diphones not in the database. Thus both methods could be synthesized from utterances from the same database by regular synthesis procedure in Festival.

For synthesis Festival 2 is used. This is currently the newest version of Festival speech synthesizer described in (Clark et al. 2004). The main feature of this version is the general purpose unit selection engine, *multisyn* which allows unit selection synthesis in unlimited domains. The synthesis process in Festival can be divided in two steps: linguistic processing and waveform generation.

The linguistic processing in Festival includes tokenization, normalization, i.e. expansion of tokens to words which have associated pronunciations in the pronunciation lexicon, POS tagging and phonetization. The result of linguistic processing is the utterance structure where all linguistic information gathered during the text processing is saved in features and organized in relations mentioned in section 4.2.2.3.

Additional information on phonetic transcription and pronunciation, stress marking and syllable structure of the words is added to the utterance structure in the phonetization. Phonetization is done by lexical look up. The pronunciation lexicon contains information on pronunciation of words, assignment of lexical stress and syllable structure. Normally, a small list of usually used words, which are not in the lexicon (so called addenda) is also consulted if a word is not in the lexicon. For words neither in lexicon nor in the addenda, letter-to-sound rules have to be applied. However, no addenda or letter-to-sound rules were used in the synthesis of the test word pairs here. Instead, a small polyglot pronunciation lexicon ”newlex” was built to contain all utterances which will later be included in the synthesis. This model solution was enough for synthesis of word pairs for purposes of the experiment, but it would not be sufficient for real world synthesis in unlimited domains. The phonetic transcription in the lexicon was based on the predefined phoneset *multiling_phones*. The phoneset was compiled out of English, German and Bosnian phonesets used in the initial transcription of the corpora. These phonesets were also used in transcription of single language corpora

as described in section 2. In the decision which phones to include in the phoneset a simplifying assumption has been made that all consonants are same phones. This is probably also true for my non-native pronunciations of English and German consonants, although in the case of /l/ some differences exist. However, this was not considered crucial for the synthesis of the test cases. All word pairs for synthesis have been recorded, and I relied on unit selection engine which would automatically select phones from the correct language because join costs are zero.

The final step in the synthesis procedure is unit selection and waveform generation. The units are selected and concatenated automatically by multisyn algorithm based on minimization of target and join costs.

This standard synthesis procedure is used to synthesize the word pairs for FULL and ”PAUSE” method. As it will be indicated below other cross-word handling methods use slightly different synthesis.

Pause insertion was done by including a pause entry in the lexicon. It’s pronunciation was set to silence. In the synthesis the word string ”<Bosnian word> <PAUSE> <English/German word>” was the input to Festival. This added an extra silence between the words in the word pair.

A look at the database for method FULL reveals that the cross-word diphone in the synthesized test word pairs and the one recorded in the database are in the similar phonetic environment. This was necessary for two reasons. First, it was important to have words without additional within-word joins. As described above, this was done in order to keep the word boundary the only potentially problematic concatenation place and thus ensure that the judged quality is actually the quality of the word boundary. Thus all words for test prompts had to be in recorded in the database. Second, same cross-word diphone had to be tested in different word pairs in order to avoid priming effects on intelligibility test. Similar environments of cross-word diphones were chosen to ensure comparability across same cross-word diphone synthesized with different methods. Thus several examples of similar environments of the same diphone were present in the database. When a diphone was synthesized in the context not in the database, the better, similar contexts were automatically chosen, because of lower join costs, although there were examples of other contexts for the same diphone. Thus the synthesized examples do not really show the really problematic cases for the FULL method, where a diphone from one context is synthesized in a fully different context,

so it can be expected that the results will be obscured by this fact. This means the FULL method will probably be perceived as clearly better method than all the rest.

Phone concatenation

The voice for implementing phone concatenation (”PHONES”) method *(multiling_phones_multisyn)* for dealing with cross-word diphones is built from the database made of recorded single words as shown in the appendix C.2. The phone concatenation algorithm is activated if no diphone is found in the database. By building a database of single words from the word pairs which should be synthesized, it is guaranteed that no cross-word diphone will be found in the database. At the same time, there will be no within-word joins, since all words for synthesis are included in the database.

As already mentioned in section 3.2.2, the phone concatenation synthesis procedure implemented in Festival extends the word final phone to the right and the word initial phone in the second word of the pair to the left on synthesis time. The word pair is thus synthesized without having the cross-word diphone in the database.

Nativization

The nativization method ”NATIVE” includes mapping of a foreign language phone (i.e. an English of a German phone) to the closest phone in the basic language (Bosnian). The most similar Bosnian phone is not found automatically, but the mapping of Bosnian phones to the foreign ones is defined for the language pairs Bosnian-English and Bosnian-German according to informal productional testing (cf. section 3.2.1). When the closest match in Bosnian is found, a diphone consisting of a Bosnian phone and the initial part of accommodated foreign phone is chosen from the inventory.

Like phone concatenation, the nativization method is also activated when a diphone is not found in the database. The back-off rules implemented in Festival are used for nativization (Clark et al. 2004). These rules replace a missing diphone by a predefined Bosnian counterpart. The list of substitution is defined in the lexicon used in the synthesis. The substitution rules are in form (x y). The substitution means that y is substituted for x, so if a diphone x_z is not found in the database, the diphone y_z is substituted for x_z if y_z is in the database.

The database for the ”NATIVE” method was designed to allow the predefined phone substitution. The voice implementing nativization *(multiling_native_multisyn)* method is built from the database given in appendix C.3. This organization of the database allows correct replacement of English and German phones with Bosnian ones. For example, assume that the word pair ”punac approve” (*father-in-law approve*) should be synthesized. Since the schwa vowel /ax/ is not a part of Bosnian phone set, it has to be replaced by a Bosnian phone. The substitution rule in the lexicon is (ax a), which means that schwa should be replaced by vowel /a/. This means that the diphone /ts a/ must be in the database, otherwise the replacement cannot be done, since the substitution is also missing. The Bosnian vowel a is thus recorded with the Bosnian word ”punac” in the word ”apel”. When synthesizing the word pair ”punac approve”, the cross-word diphone /ts ax/ cannot be found, so the substitution takes place and substitutes ax for a. The phone following the substituted diphone is not changed to keep the spectral continuity as also noted in (Clark et al. 2004). The resulting diphone series is thus /ts a/ /ax p/. Figure 4.2 shows that both the vowel /a/ and /ax/ are realized in the utterance. However, this doesn’t seem to be audible and problematic in this case.

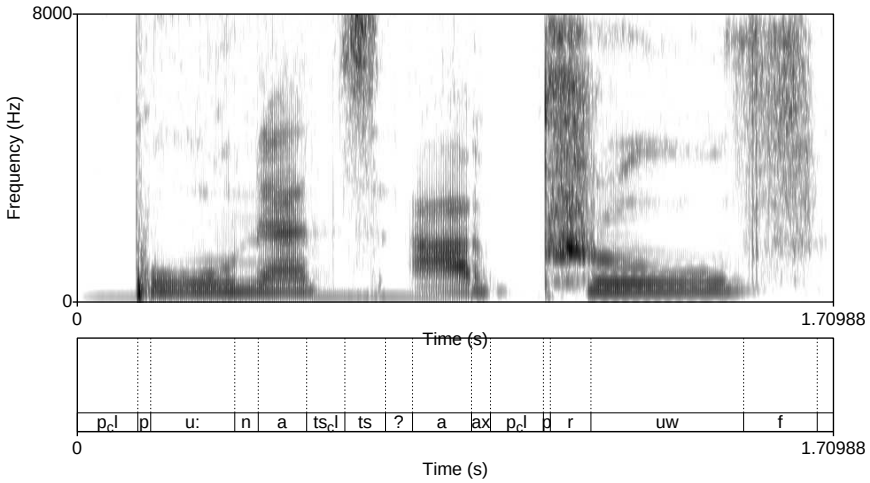

Figure 4.2: Spectrogram of the word pair *”punac approve” (father-in-law approve)* indicates that both the vowel /a/ and schwa vowel /ax/ are realized in the utterance because the back-off rules do not accommodate the right context of the substituted phone

4.2.4 Experimental design

Four approaches to handling multilingual cross-word diphones are compared on the basis of intelligibility and naturalness of speech resulting from use of one of these methods in the synthesis.

Subjects were speakers of Bosnian and English, Bosnian and German or Bosnian and both foreign languages. The experiment was open to subjects reasonably fluent in Bosnian and either English or German. It would be difficult to find subjects native in both Bosnian and English or German. However, it was assumed that in intelligibility assessment persons reasonably fluent in both languages can perform as well as native speakers. The notion of naturalness is here defined as spectral quality of the mutlilingual cross-word boundary, and the expectation is that the spectral quality of the sound can be judged objectively by both natives and non-natives. To confirm or reject this a comparison between native and non-native judgements in all languages would be needed. However, the expectation was that most of the subjects will be native speakers of Bosnian with good command of the foreign language, so that such comparison will not be possible.

34 subjects were involved in the experiment for Bosnian-English word pairs. For Bosnian-German, 30 subjects took part in the experiment. Four subjects did the experiment for both language pairs, so their results were not considered in the analysis of Bosnian-English word pairs. Thus there are two non-intersecting groups of 30 subjects for both language pairs. All subjects were native speakers of Bosnian fluent in a foreign language. The range of subjects' ages was from 17 to 52, most subjects were at the age of 22.

The experiment had to be conducted on the internet since no subjects fluent in Bosnian and either foreign language could be found (*http://www.ling.ed.ac.uk/ s0343746/exp.html*). The subjects had to judge 20 word pairs for their intelligibility and naturalness. The experimental hypothesis was that subjects' understanding and naturalness judgments of the speech synthesized from database including one example of a cross-word diphone for each langauge pair (FULL method) are same or worse than understanding and naturalness judgments of the speech synthesized by one of other three methods described in previous sections. This is tested against the null hypothesis that differences in intelligibility and naturalness as perceived by subjects are due to random variation rather than to different methods for handling cross-word diphones.

4.2.4.1 Intelligibility

In the intelligibility test the subjects were asked to listen to each word pair not more often that twice and write down what they heard. It was expected that intelligibility performance will be high due to high ability of humans to guess the correct word even if they are not exactly sure what they heard. This normalization effect is partly smoothed by the fact that nonsense word pairs are listened to, so a sentence context cannot be used to guess the word. However, humans possibly use also other cues to determine what they heard which are perhaps not removed by having no semantical context. In any case, in order to get a clearer idea about how good the intelligibility of word pairs synthesized by different methods is, it is helpful to see how good people perform on natural speech. The results on natural speech serve as reference value, against which the intelligibility of synthesized word pairs is compared. In order to set this reference value, 4 recorded speech word pairs were inserted as control, one word pair for each test diphone in each language pair. In the intelligibility test the subjects recognized 104 Bosnian-English and 111 Bosnian-German recorded word pairs correctly. For both language pairs the total number of word pairs was 120, so 86.6% Bosnian-English and 92.5% Bosnian-German word pairs have been recognized correctly.

4.2.4.2 Naturalness

Two experimental methods for judgments of naturalness are often used in evaluation of synthetic speech. The first method is to ask subjects to listen to the speech and rate it. Rating can be done according to a scale set by the experimenter. Another kind of rating task is so called magnitude estimation experiment. The subjects are presented first with a synthetic utterance, called standard stimulus and asked to assign it a freely chosen order of magnitude (Sorace 2003). Each following utterance should then be rated relative to this standard stimulus. The second evaluation method is so called forced choice experiment. In this kind of experiment subjects are presented with pairs of utterances where same utterance has been synthesized by different synthesis methods to test. They are asked to choose the better sounding synthetic utterance.

For testing purposes in this project all three experimental methods for assessing naturalness have advantages and disadvantages. If a rating scale is predefined, the subjects have to be able to keep in mind how the speech they hear is allocated to different levels of the scale. For example, if a word pair is given a middle mark on the scale,

this marking should be taken into consideration when rating the next utterance. If the forced scale is too fine grained the subjects might make arbitrary differences, which results in higher probability that ratings are random. On the other hand, if the scale does not offer enough rating points, people would not be able to express differences they eventually hear.

These problems are not present in the magnitude estimation experiment, where subjects can set the scale on their own, assign some mark to the first word pair they hear and rate each following word pair relative to the first one. It is also an advantage to have naturally set scales since this gives the subjects freedom to express as many distinctions as they can make. Therefore it can be assumed that every person will be able to judge more accurately on a self-defined scale than using some other forced scale.

As noted in (Bard et al. 1996) a problem of magnitude estimation in linguistic application is that there is no objective physical quantity that linguistic judgments can be compared to. This also applies to synthesized speech. Whereas in the case of line length for example, the estimated line length can be compared to the real line length, non such objective measure exists for speech. Bard et al. (1996) suggest so called cross-modal matching to solve this problem. In a magnitude estimation experiment the subjects should rate both the objective measure, line length and linguistic stimuli. If the ratings for both correlate, the magnitude estimation judgements can be validated. Judging the line length can be understood as practice or training for the actual experiment, so every magnitude estimation experiment generally requires training.

Another problem of magnitude estimation is that people often choose not to use a natural scale but keep to some standardized scale instead, usually school marks. Bard et al. (1996) and Sorace (2003) report this tendency of choosing the scales. If this is the case, the advantage of freely set scales and using the full range of differentiation possibilities is not used, so the justification for actually doing magnitude estimation experiment instead of predefined scales disappears. It is possible to prevent subjects from using standardized scales by instructing them explicitly not to do so.

Forced choice is generally an easier task for subjects. One reason is that no training is needed. In the cases where there is only a subtle difference between two word pairs one of them has to be chosen, so the decision what to do is easier. However, this is not necessarily the advantage from the experimenters' point of view since this might introduce higher factor of chance if subjects just choose any of the two word pairs.

Given these advantages and disadvantages of experimental methods, the decision was made to employ both methods for naturalness assessment. The aim was to compare the naturalness judgments in two naturalness tests and see whether the judgments differ when humans are forced to choose between two synthetic word pairs compared to the case where they have freedom to make their own judgment.

For judging naturalness the subjects first performed magnitude estimation test. They were asked to judge the quality of the synthesized word pairs they heard. Subjects were also instructed to pay attention only to the quality of the sound, and not to score whether word pair makes sense, or whether it is appropriate for certain context, for example. This was done in the attempt to focus the subjects' attention to spectral quality of the word boundary and exclude other possible effects on their judgements, which would confound results. A scale for judgments has not been suggested. The subjects were free to select their own scale, but they were asked to judge relative to the reference word pair (standard stimulus). The reference word pair was a natural speech utterance. The comparison relative to the natural speech reference seemed a good way to estimate which method for handling cross-word diphones results in synthetic speech closest to the recorded speech.

Subjects were not warned against using short or standardized scales. However the example of rating showed a rating of 65.5 which influenced several subjects to take a larger scale. 7 subject chose percentage scale from 1 to 100. Short scales from 1 to 10, or 1 to 5, which are common marking scales in schools and at the university in Bosnia, were chosen by 21 subjects. 32 subjects chose a scale different from these two. This confirms the observation that subjects like to use some standardized scales (in this case percentages or school marking scales), but they also can be influenced to set an individual scale.

An example of judging relative to a reference was given using the example of the line length judgments found on the internet (Corley et al. 2004). The length of the line was used as the example. However, in order not to prolong the time needed for experiment and thus the subjects' readiness to do it, the line length was not used as control condition. Subjects were not asked to perform the rating of the length of the line, so cross-modal matching cannot be done here.

In addition to magnitude estimation test, forced choice test was conducted. In the forced choice part of the experiment, the word pair synthesized by FULL method

was compared, for each diphone, to the synthesis of the same diphone by a different method. The order of presentation for pairs of word pairs was randomized, and the order of presentation of word pairs by FULL and another method within the word pair was also varied. This should exclude the possibility that the order of presentation of word pairs affects subjects' ratings.

4.3 Results and Discussion

4.3.1 Intelligibility

In the intelligibility experiment the number of correctly recognized word pairs for each method and each language pair was counted. The decision whether an answer is correct depended on the orthography. Not all subjects were good in writing, both Bosnian and a foreign language, so their answers might be orthographically wrong although they recognized the word correctly. If a word pair is written in wrong orthography, it was nevertheless declared correct if there was no ambiguity in what a person could have heard. For example if the English word *append* is written *apend* this was recognized as correct. On the other hand, however, if in the Bosnian-German word pair ”sinoć Pflüge” (*last night, ploughs*), German affricate /pf/ is written as ”F” rendering the word Flüge (*flights*), the word pair was declared wrong.

A further distinction was made between correct recognition of the whole word pair and correct recognition of the word boundary. In the Bosnian-English word pair ”kašalj their” (cough their) the English word ”their” was often recognized incorrectly (probably due to my peculiar pronunciation) as ”air” or ”layer”. Several subjects, however, identified the word ”there”. In the first case, it is clear that the cross-word diphone was misunderstood and the whole word was recognized incorrectly. In the second case, the word boundary is recognized correctly, however, the recognized word was wrong. The intelligibility results are given separately for overall recognition correctness of the word pair, where a word pair is wrong when either word is written wrongly, and correctness of recognition of the word boundary. In the second case, word pairs are declared as correct if the word boundary is recognized correctly, disregarding the correctness of the rest of the words.

The results for overall intelligibility of word pairs are given in figure 4.3. Binomial test

shows that the null hypothesis that word pairs are guessed by chance can be rejected ($p<.05$) for each method in both language pairs.

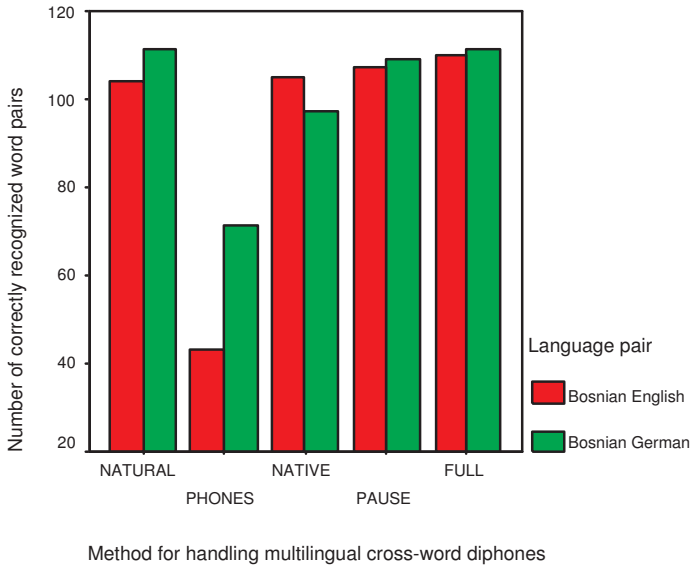

Figure 4.3: Number of correctly recognized word pairs when both words are recognized correctly in the intelligibility experiment (Total: 120)

The results show that the highest number of correctly recognized word pairs was achieved when word pairs were synthesized by FULL method. The best intelligibility among alternative methods is pause insertion, followed by nativization. Word pairs synthesized by phone concatenation had the lowest number of correct recognitions.

These results are consistent for both language pairs. However, there are clear differences in distributions of correct and incorrect answers within methods for two language pairs. In Bosnian-English word pairs all three methods, FULL, PAUSE and NATIVE exceed the topline of 86.7% correctly recognized recorded word pairs. In Bosnian-German case the number of correctly recognized word pairs in FULL method achieved the topline of 92.5%, but it does not exceed it as in case of Bosnian-English word pairs. 90.8% of word pairs synthesized by phone insertion are recognized correctly which is also close to the topline. The percentage of 80.8% correctly recognized nativization examples is substantially lower than FULL and PAUSE recognition rate. For both language pars phone concatenation resulted in lowest number of correctly recognized word pairs. Only 35.8% of Bosnian-English word pairs synthesized by phone concatenation were recognized correctly. For Bosnian-German the percentage of correctly recognized phone concatenation word pairs is 59.2%. Although more

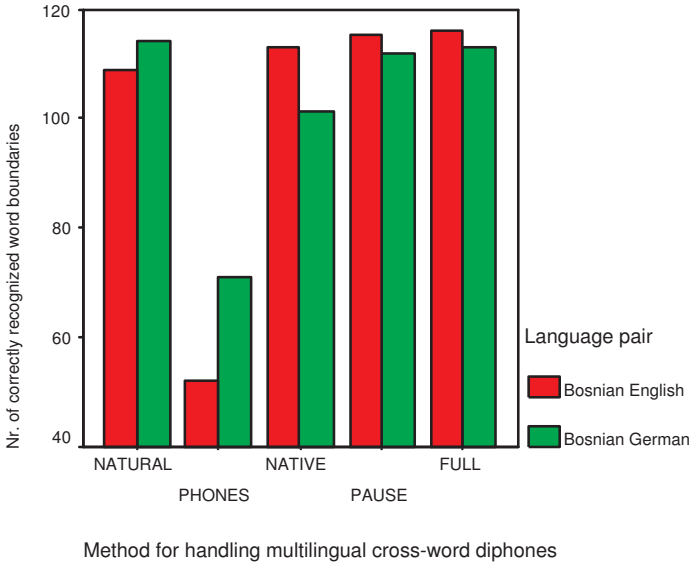

Figure 4.4: Number of correctly recognized word pairs where word boundary is recognized correctly in the intelligibility experiment (Total: 120)

Bosnian-German than Bosnian-English phone concatenation word pairs are recognized correctly, the percentage of 59.2% is still substantially lower than recognition rate for other synthesis methods.

If the correctness of the word boundary is considered the results change slightly. Naturally, the number of correct answers increases since word pairs with correct word boundary are added to all fully correct answers. Figure 4.4 shows the results.

The percentage of correct answers for all methods in overall recognition is 78.2% for Bosnian-English and 83.2% for Bosnian-German word pairs. The percentage of correct answers in word boundary recognition increases to 84.2% for Bosnian-English and 85.2% for Bosnian-German word pairs. The increase in number of correct answers happens both for recorded word pairs and word pairs synthesized by all methods. Thus although the increase for all methods might appear high, especially in Bosnian-English case, the overall tendencies in results do not change. The FULL word pairs still have the highest recognition rate, followed by PAUSE and NATIVE and the recognition rate of PHONES method is the lowest.

In Bosnian-German the FULL, PAUSE and NATIVE word pairs are again below the topline of 95%. The difference between correctly recognized Bosnian-German word boundaries in recorded speech and FULL method is only one word pair, between

recorded speech and PAUSE it is 2 and for NATIVE method 13 word pairs. These differences are same as in the case of overall word recognition, except for FULL method. Thus the proportion in the number of correctly recognized word pairs does not change. The number of correctly recognized PHONES word pairs remains the same for Bosnian-German word boundary recognition. Thus in all incorrectly recognized PHONES word pairs the word boundary was also not recognized correctly. This indicates that the PHONES method is indeed inferior to the other methods considering intelligibility of the word boundary in Bosnian-German examples.

This observation cannot be entirely confirmed for Bosnian-English word pairs. There, the number of correctly recognized word boundaries in PHONES method increased from 35.8% to 43.3%. However, relative to the other methods this is still low recognition rate. The recognition for Bosnian-English is still above the topline for FULL, PAUSE and NATIVE word pairs. The difference in number of correctly recognized word pairs between two best methods, FULL and PAUSE is only one word pair and between FULL and NATIVE three word pairs.

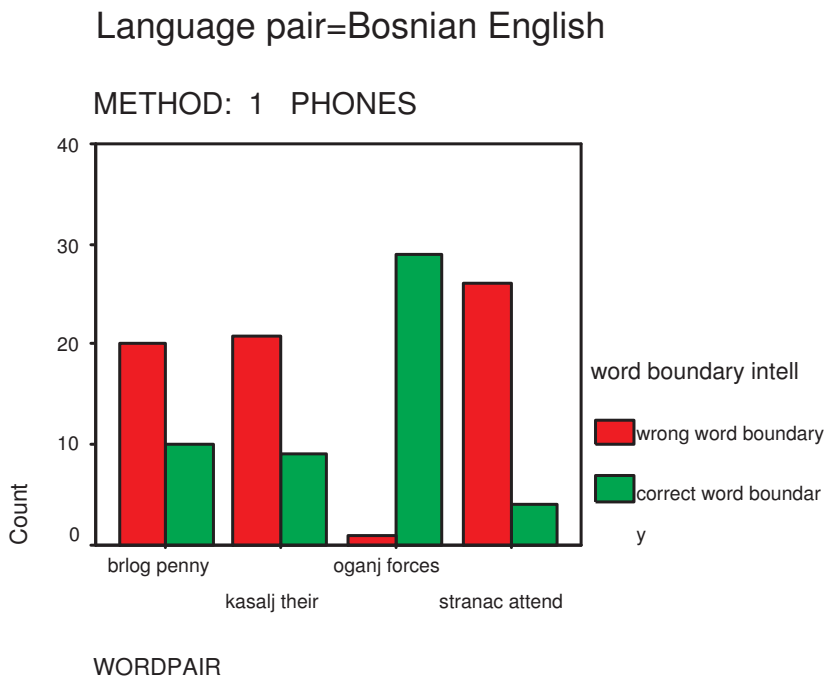

Figure 4.5: Word boundary recognition rate for Bosnian-English word pairs synthesized by PHONE method

These constant results between overall and word boundary recognition indicate that in

most cases where a word pair was incorrectly recognized the word boundary was also incorrect. This was the expected result if the word boundary realization affects overall recognition of the word pair.

Low recognition rate of word pairs synthesized by PHONE method requires some explanation. The recognition rates of single word pairs is investigated in order to see which word pairs were particularly problematic for the method. Figure 4.5 gives the word boundary recognition rates for Bosnian-English word pairs. It shows that the word boundary in the word pair ”oganj forces” (*fire forces*) has been recognized correctly by 29 out of 30 subjects. Three remaining word pairs, on the contrary, have very high rate of wrong word boundary recognitions. It is possible that the results for the word pair ”kašalj their” (*cough their*) are influenced by my peculiar pronunciation of the word ”their”. However, in the other two word pairs the results can be attributed to the PAUSE method. The problem with the word pair ”stranac attend” (*foreigner attend*) is that the release for the word final affricate /ts/ is missing. Figure 4.6 shows spectrum of the word pair when synthesized by phone concatenation.

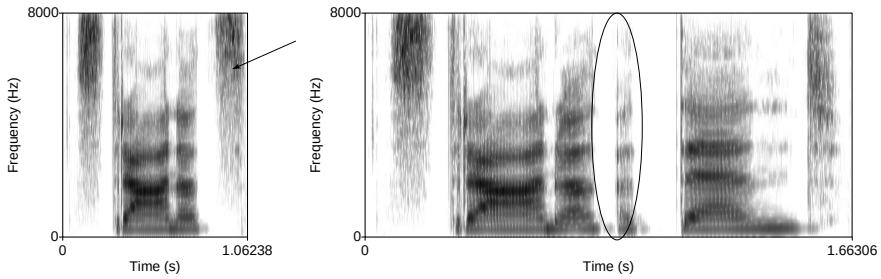

Figure 4.6: The spectrogram of the word pair *”stranac attend”* shows that fricative /s/ in the word final affricate /ts/ is missing when phone concatenation is used

The spectrogram of the word ”stranac” is shown on the left. It indicates clearly the fricative part in the word final affricate. On the right hand side the word pair ”stranac attend” is shown. The extension of the schwa vowel seems to cover the fricative part of the affricate, so only the closure can be heard. The affricate thus cannot be heard and the word pair sounds like ”strana attend”. ”strana” is a word of Bosnian meaning ”page” or ”side”, so in the intelligibility test 26 out of 30 subjects could not recognized the word boundary pair properly. The same problem is present in the word pair ”brlog penny” (*mud penny*). The word final stop /g/ is not realized, so the word pair is in most cases recognized as ”vrlo penny” (*very much penny*) or ”grlo penny” (*throat penny*), which shows the tendency of people to normalize what they heard to some existing

word in the language. Leaving out of sounds at the word boundary in PHONES word pairs could be attributed to the wrong labelling, however, the labels for these word pairs were manually corrected. Another explanation is that there is a bug in the phone concatenation algorithm which sometimes causes parts of sounds not to be concatenated properly.

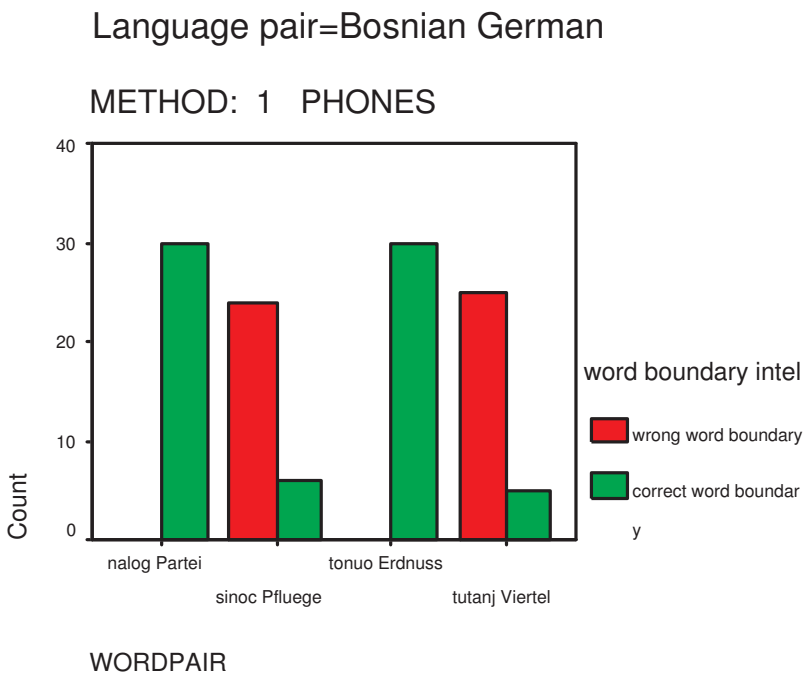

Figure 4.7: Word boundary recognition rate for Bosnian-German word pairs synthesized by PHONE method

As figure 4.4 indicates, word boundary recognition of Bosnian-German word pairs synthesized by phone concatenation is higher than that of English, however, still low compared to other methods of synthesis in German. Figure 4.7 shows that both word pairs ”nalog Partei” (*order party*) and ”tonuo Erdnuss” (*he sank peanut*) are recognized correctly by all subjects. The problematic cases are ”sinoć Pflüge” (*last night ploughs*) and ”tutanj Viertel” (*roar(n.) quarter*). The problem with the first word pair was that it was recognized as ”sinoć Flüge” (*last night flights*). Spectrogram of the word pairs in figure 4.8 shows that /pf/ closure is realized, so it cannot be assumed that the stop has been cut off, as in the English examples above. The reason for the recognition of German affricate /pf/ as /f/ might be attributed to the higher frequency of the word ”Flüge” in every day usage and my pronunciation of the affricate.

In the second word pair, the problem was wrong labelling. The spectrogram in figure

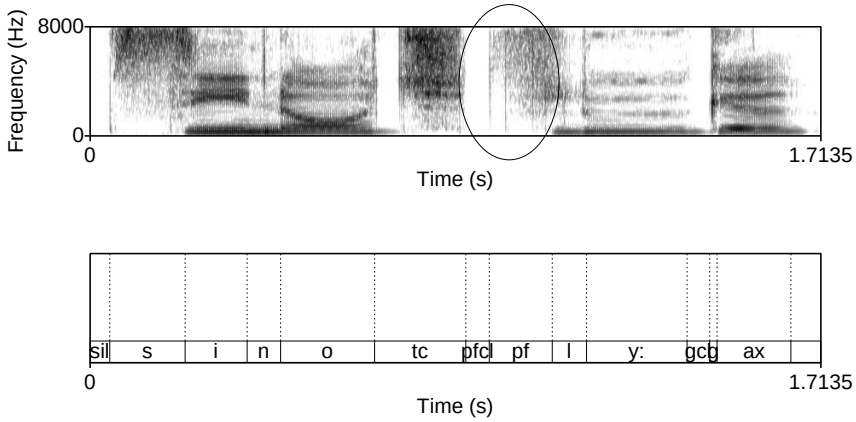

Figure 4.8: Spectrogram of Bosnian-German word pair ”sinoć Pflüge” (*last night ploughs*)

4.9 shows that there is additional sound at the word boundary. Although labelling for the test word pairs was generally corrected manually in the problematic cases, it was obviously left out here. Labelling is a problem for PHONES method as pointed out in section 3.2.2, however, it is not sure how it would have affected other methods if it was not corrected manually. Thus for German, low recognition rates for PHONES method, might also be due to factors other than synthesis method.

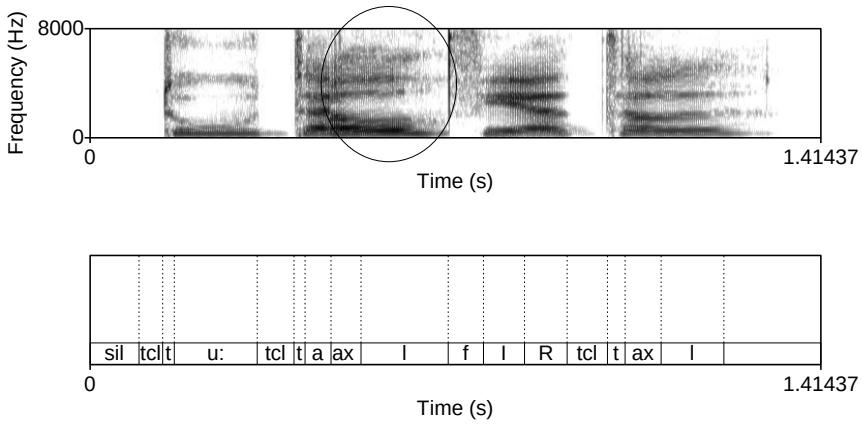

Figure 4.9: Spectrogram of Bosnian-German word pair ”tutanj Viertel” (*roar(n.) quarter*) shows bad labelling

It is also possible that understanding skills in both Bosnian and one of the foreign languages affects overall recognition. In the experiment subjects were asked to characterize their understanding and writing skills in both Bosnian and both foreign languages as ”excellent”, ”good” or ”not so good”. Table 4.1 shows ratings of language skills of

the subjects.

Language skill	*Rating*	*Percentage of subjects (%)*		
		Bosnian	*English*	*German*
Understanding	excellent	95	66.7	65.4
	good	5	33.3	30.8
	not so good	0	0	3.8
Writing	excellent	93.3	60	63.3
	good	6.7	36.7	26.7
	not so good	0	3.3	10

Table 4.1: Writing and understanding skills of subjects in the experiments

It was investigated whether there was a significant association between the language skills and recognition rates. The hypothesis was that people who estimate their writing or understanding skills lower will perform worse on overall word pair recognition and word boundary recognition. This hypothesis could not be retained in Chi-Square test on significance level 0.05 for writing or understanding ability of Bosnian and English. Neither overall recognition rates, nor word boundary recognition rates showed significant relationship with language skills for these two languages. For German understanding and writing there was significant difference in scores between three levels of German language skills. Both overall and word boundary recognition rates show significant association with German writing and understanding skills. This means that results for German overall recognition the are also due to language skills of the subjects and cannot be entirely attributed to the method of handling multilingual cross-word diphones.

4.3.2 Naturalness

4.3.2.1 Magnitude Estimation

Since every subject set his own scale for magnitude estimations, the results had to be normalized across subjects in order to attain comparable results. For every subject coefficients have been calculated by dividing each subjects' magnitude estimation for a word pair by the estimation of the standard stimulus (i.e. recorded speech). This

created a common scale for all subjects. The results were than turned into their decadic logarithm values for obtaining normal distribution (Bard et al. 1996). Further analyses were performed on data transformed in this way.

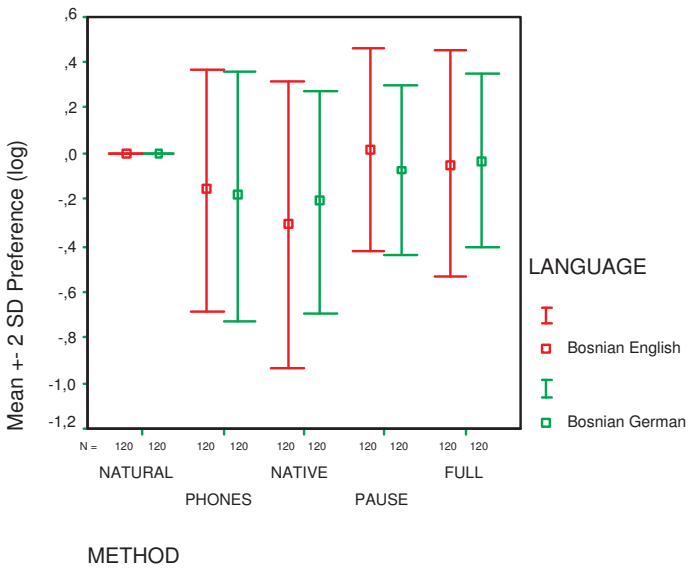

Figure 4.10: Means and standard deviations of normalized, log transformed magnitude estimates

Mean ratings along with standard deviations for each method are shown in figure 4.10 for both language pairs. The mean magnitude estimate of Bosnian-German word pairs synthesized by FULL method is closest to the mean of the recorded speech which served as reference. Among alternative methods, pause insertion is closest to natural speech, and it is also close to the preference mean of the FULL method for Bosnian-German. Word pairs synthesized by phone insertion and nativization are judged furthest from the recorded speech. Results differ for Bosnian-English. There, the most preferred method is pause insertion, which is very close to recorded speech. It is followed by FULL method. As for Bosnian-German PHONES and NATIVE have lowest preferences compared to recorded speech. The differences in means between methods are larger for Bosnian-English than for Bosnian-German word pairs.

These initial observations suggest that there are differences in human listeners' preferences of the speech synthesized from databases with different methods for handling multilingual cross-word diphones. The significance of the differences was tested by one-way ANOVA. In our experimental design, all subjects were involved in all experimental conditions (i.e. all methods for handling multilingual cross-word diphones),

so repeated measures ANOVA was chosen. The most interesting question to test is whether the difference in preference between PAUSE and FULL method and between both these methods and recorded speech for both language pairs is significant. Insignificant difference between either method and recorded speech would mean that speech synthesized by the method is perceived as close to recorded speech. Significant difference between PAUSE and FULL would mean that quality of word pairs synthesized by inserting pause between words is perceived as same (for Bosnian-German word pairs) or higher (for Bosnian-English word pairs) compared to the speech synthesized from the databases including multilingual cross-word diphones. In terms of reducing the size of the database, this would be encouraging results, since same or better speech quality can be achieved by a method which is easy to implement and saves including multilingual units in the database.

The results of the one-way ANOVA for within-subjects effects showed significant main effect of the method on subjects' quality judgements for both language pairs ($F=30.890$, $p<.001$). This means that both for Bosnian-English and Bosnian-German there is significant difference in magnitude estimations for word pairs synthesized by different cross-word diphone handling methods. The question of interest is however, whether there are significant differences in ratings between single methods, so post hoc test was conducted to compare methods pairwise.

Pairwise comparisons for Bosnian-English showed significant difference between judgment of recorded speech and preference judgments for both phone insertion and nativization ($p<.001$). Also FULL and PAUSE methods are significantly better preferred than these two methods ($p<.001$). Nativization is significantly less preferred method of the two ($p<.001$). Figure 4.10 showed that PAUSE is the method which is rated closest to the recorded speech. Significance test confirms this. The difference in preference rating between PAUSE and recorded speech is insignificant ($p=.976$) and so is the difference between FULL and recorded speech ($p=.615$). There is no significant perceived difference between PAUSE and FULL method ($p=.259$). Thus the perceived quality word pair synthesis by pause insertion is at least same as that of inclusion of a multilingual diphone in the database and also very close to the perceived quality of the recorded speech.

As figure 4.10 showed, for Bosnian-German word pairs, FULL method is the one with preference judgments closest to the recorded speech. The difference in magnitude estimates between recorded and FULL word pairs is not significant ($p=.815$). Inserting

pause also does not differ significantly from the recorded speech ($p=.055$), but the insignificance of the difference is not persuasive. However, since the FULL is not significantly preferred over PAUSE ($p=.490$) it can be concluded that also for Bosnian-German word pairs the quality of synthesis for inserting pause is at least as good as that of including a cross-word diphone in the database. Phone insertion and nativization word pairs are significantly less preferred than any other cross-word diphone handling method and recorded speech ($p<.001$). Unlike for Bosnian English word pairs there is no significant difference in perceived quality between these two methods ($p=.896$).

These results suggest that we can retain the initial experimental hypothesis that inclusion of multilingual cross-word diphones in the database is not necessary. Although phone concatenation and nativization were not highly rated, pause insertion was at least as good in perceived synthesis quality as full coverage of multilingual diphones. It should also be considered that only good examples of full coverage were tested, as described in section 4.3.1, so the quality of the tested word pairs sets a topline for what the method can achieve. Thus if pause insertion is at least as good in the synthesis of the test set, it can be expected to be superior to inclusion of one example of diphone in the database in real world synthesis, where coverage problems for FULL method affect synthesis quality.

It remains uncertain how reliable the subjects were in doing magnitude estimation. Normally, a correlation with line length judgments would be used as control condition. For line length it has been shown that it is proportional to the actual line length. So it can be assumed that if the subjects can judge line length reliably, they also can judge speech. As already mentioned, in this experiment subjects were given the line length as an example, but they were not asked to judge the line length. So it cannot be measured how reliable their judgments are.

4.3.2.2 Forced Choice

The results of the forced choice experiment are given in Figure 4.11 and table 4.2. The number of preferred word pairs sorted by method for handling multilingual cross-word diphones are shown for each language pair.

The overall results show that subjects clearly prefer synthetic word pairs resulting from the inclusion of one example of a multilingual diphone in the database (FULL) over word pairs synthesized by alternative methods (NATIVE, PAUSE and PHONES). Bi-

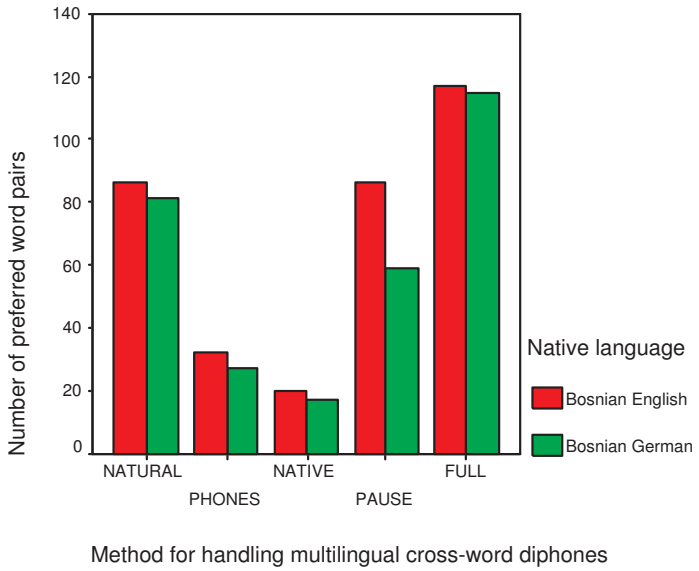

Figure 4.11: Number of preferred word pairs in forced choice experiment for each cross-word diphones handling method and each language pair. Total number of word pairs in each condition is 120.

nomial tests were carried out for each method separately to test the hypothesis that there is significant preference between word pairs within that method. The null hypothesis was that subjects have no significant preference for word pairs synthesized by a particular method, but the preference ratings are assigned by chance. The null hypothesis could be rejected for Bosnian-English word pairs synthesized by all methods ($p<0.05$). Among Bosnian-German word pairs, the null hypothesis could be rejected ($p<0.05$) for all methods except pause insertion. The probability that rejecting the null hypothesis of random preference rating for PAUSE word pairs is wrong is $p<0.927$. Thus there is possibility that ratings occur randomly.

The surprising result is that the overall preference for the word pairs synthesized by method FULL is also more frequent than preference for recorded word pairs (NATURAL). The expected situation is that people would prefer recorded to synthesized speech.

There are two possible explanations for the high number of preferences for the FULL method. The first reason might be that only good examples of FULL method word pairs are included in the database as explained in section 4.2.3. This suggests that in cases where a diphone is covered in the database the quality of resulting synthesis is superior to synthesis of cross-word diphone by any other alternative method. However,

Method		*Number of preferred word pairs*	
Nr.	Example	*Bosnian - English*	*Bosnian - German*
1	FULL	117	115
2	NATURAL	86	81
3	PAUSE	86	59
4	PHONES	32	27
5	NATIVE	20	17

Table 4.2: Number of preferred word pairs in forced choice experiment for each cross-word diphones handling method and each language pair. Total number of word pairs in each condition is 120.

this does not explain higher preference for FULL method over recorded speech which is also of a very good quality.

A possible confounding factor in the forced choice is high frequency of word pairs synthesized by FULL method. In each pair of word pairs a word pair synthesized by FULL method is compared to a word pair containing same diphone and is synthesized by an alternative method. Thus FULL word pairs were often repeated in the experiment (more precisely, four times for each test diphone). The frequency of these word pairs might have affected subjects' naturalness ratings. The alternative would have been to take a different word pair for each comparison in the forced choice test. Not enough words could be found which fit into general requirements for test words as explained in section 4.2.1. Taking any words would introduce the possibility that preference is due to word pair rather than to the method. Thus to keep the comparability between methods, the same FULL word pair was used in all forced choice comparisons and the risk of unwanted effect of frequency on preferences was tolerated. If frequency does not affect the preferences, the results suggest that quality of the FULL multilingual word pairs is very close to the original recorded speech, and differences in quality are often not perceivable.

Among the three alternative methods, pause insertion seems to have the best quality. For Bosnian-English word pairs the number of preferences for PAUSE method is same as number for preferences for recorded speech, so differences between recoded word pairs and word pairs synthesized by PAUSE method were not at all perceivable. For Bosnian-German the difference between number of preferences for PAUSE and the

FULL method is substantially larger than the difference in preference counts between these two methods for Bosnian-English. However, given that the number of preferences for PAUSE method in Bosnian-German is not higher than it would be by chance, this result is not reliable.

Unlike magnitude estimation results, the results of the forced choice experiment confirm that it is not straightforward to retain the original experimental hypothesis that any of the three alternative methods would be preferred to the inclusion of the diphone in the database. As in magnitude estimation experiment, inserting a pause was the most preferred method among alternative methods. In the forced choice experiment it even reached the same preference rate as natural speech for Bosnian-English word pairs. For Bosnian-German word pairs pause insertion was also the best alternative method, however, substantially less frequently preferred than FULL method. The preference results for PAUSE method, however, should be taken with caution since the hypothesis that they are random could not be rejected at significance level 0.05. Although preference results for pause insertion in forced choice are good, they are not better than full coverage of diphones in the database. However, this also might be due to the frequency effect. The results of the two naturalness tests have same tendencies. Both naturalness tests show that full nativization is least preferred method. Both test also indicate that FULL method is the best method, followed by PAUSE. In the forced choice test, however, it could not be tested how close these two best methods are. Magnitude estimation experiment confirmed that if people are free to make fine decisions about the quality of speech, they would not make difference between FULL and PAUSE. Both these methods are also very close to recorded natural speech.

4.4 Summary

In this chapter four methods of handling diphones at the boundaries of words from different languages have been tested. Full coverage of one example of each multi-language cross-word diphone in the database served as baseline. The goal of the experiment was to examine whether alternative methods involving only single language diphone coverage can produce at least same quality synthetic speech. Bosnian-German and Bosnian-English word pairs were used for testing. The diphones tested were representative of selected potentially problematic and potentially unproblematic cross-language word boundary diphones. Three alternative methods: full nativization of for-

eign sounds to Bosnian, backing-off to phone concatenation where no diphone is found and inserting a pause between two words from two different languages were tested on all diphones. The results suggest that including a cross-word diphone in the database results in better intelligible speech. The results of naturalness tests are not clearly cut. Magnitude estimations for pause insertion and FULL were significantly better than all other methods, but there was no significant difference between these two methods. In forced choice, pause insertion was preferred often, but not as frequently as FULL method.

Chapter 5

Conclusions and Future Work

Attaining a good coverage of units in the database is a well known problem for database design for unit selection speech synthesis, especially in unlimited domains. At the same time, the coverage of units has a direct impact on the quality of output speech since concatenation of units not in the database usually sounds bad.

In this project the coverage possibilities were investigated for diphone sized units and a polyglot database containing diphones for Bosnian, English and German. It has been shown that in polyglot databases the coverage problem is even more acute, since not only single language diphones have to be covered, but also the concatenation points of words from different languages have to be accounted for. The coverage investigations suggested that it is more reasonable to cover only single language units in the database and handle multilingual cross-word diphones on synthesis time. Three alternative methods have been suggested: resorting to phone concatenation when a multilingual cross-word diphone is encountered, nativization of a foreign phone (English or German) to a basic language (Bosnian) phone and inserting a pause between two words from different languages.

The perception tests showed that for Bosnian-English and Bosnian-German word pairs used in the experiment, the intelligibility of synthetic speech is generally very high, except for the PHONES method. Naturalness tests revealed that covering a multilingual cross-word diphone in the database and pause insertion are clearly superior to any other method. Both these methods are very close to the quality of speech when it is only recorded and played back. The best method among alternative methods was pause insertion. In the magnitude estimation test finer grained differences in quality

of speech produced by different methods could be expressed. The results showed that there is no significant difference in quality between coverage of cross-word diphones in the database and pause insertion between words on synthesis time. Thus the overall experimental hypothesis that at least one alternative method is same or better in quality of output speech than FULL method could be retained.

Given the fact that the test set included only good examples for FULL method, it can be assumed that in synthesis of unrestricted input PAUSE method would be the superior one. A further advantage of pause insertion is that it is technically the easiest method to implement, since it can be applied in any synthesizer and does not require any new extensions to the existing synthesizer. The good quality of speech synthesized by pause insertion might be explained with the tendency of people to expect a short break between words and not perceive it as disturbance.

Naturally, the experimental results presented here are only valid for synthesis of the limited number of diphones and word pairs selected here for testing. However, testing for all diphones is impossible. Although it might be possible, extensive testing of combinations of different phone groups (fricatives, vowels, etc.) requires a lot of resources. So, the results achieved here could be used as reasonable starting point for database design decisions in polyglot speech synthesis.

Building a real polyglot Bosnian-English-German unit selection voice for unlimited domain for Festival remains for future work. This task requires first extending Festival for Bosnian. The most important task here is to write a set of letter-to-sound rules for Bosnian and find a way to deal with word-accents. The latter is not easy and is probably only possible by using a pronunciation lexicon which does not exist at present. Next, a polyglot database has to be designed. Although in this project a single database containing diphones from all three languages has been constructed, it is clear that for a real unlimited domain voice this is not feasible. Good coverage of context-dependent diphones, needed for a good voice in a single language, already requires prohibitively large databases. For an arbitrary number of languages, covering all languages' diphones in the database is clearly impossible. A better solution would be to adapt the synthesizer to use several single language databases at the same time. When a foreign word is encountered in the input, the system should synthesize it using the units from the database for that particular language. The multilingual cross-word units could be handled by one of the alternative methods investigated in this project. Since inserting the pause resulted in good quality synthesis for word pairs, it could be

used rather than other two methods.

Given that most synthesizers are multilingual and already have databases for different languages, having single langauge databases with good unit coverage and handling multilingual units by pause insertion would not require any further extensions to the systems. Of course, the synthesizer has to have the possibility to switch between databases at synthesis time. In Festival, such possibility does not exist at present.

Appendix A

Table of symbols

Phone labe	*IPA*	*Phone label*	*IPA*	*Phone label*	*IPA*	*Phone label*	*IPA*
a	a	Z	ʒ	N	ŋ	n	n
a:	a:	f	f	eh	e	I	ɪ
e	e	v	v	er	ɾ	y:	y:
e:	e:	tS	t͡ʃ	ey	ei	y	y
i	i	tc	t͡ɕ	el	l̩	Y	ʏ
i:	i:	ts	t͡s	em	m̩	E:	ɛ:
o	o	dZ	d͡ʒ	en	n̩	E	ɛ
o:	o:	dc	d͡ʑ	hh	h	9	œ
u	u	l	l	ih	ɪ	U	u
u:	u:	L	ʎ	iy	i	O	ɔ
p	p	r	r	ow	oʊ	aI	aɪ
t	t	J	ɲ	oy	oj	OY	oʏ
k	k	aa	a:	th	θ	aU	au
b	b	ae	æ	uh	ʊ	a^	a̯
d	d	ah	ʌ	uw	u	E^	ɛ̯
g	g	ax	ə	w	ʋ	o^	ɔ̯
s	s	ay	aɪ	R	χ	C	ç
z	z	ao	ɔ	zh	ʒ	h	h
S	ʃ	aw	au	6	ɐ	pf	p͡f
m	m	dh	ð	2:	ø:	?	ʔ

Figure A.1: Phoneset for polyglot test voices with approximate IPA symbols of phone labels. Long vowels are marked by ":".

Appendix B

Test word pairs

B.0.1 Bosnian - English

Diphone	*Word Pair*	*Method*
g p	talog people (*sediment people*)	FULL
	razlog party (*reason party*)	NATIVE
	brlog penny(*mud penny*)	PHONES
	nalog-pages (*order pages*)	PAUSE
	prošlog-programme (*last programme*)	RECORDED SPEECH
J f	toranj forward (*tower forward*)	FULL
	bubanj forty (*drum forty*)	NATIVE
	oganj forces (*fire forces*)	PHONES
	šušanj formal (*rustle formal*)	PAUSE
	pladanj finding (*tray finding*)	RECORDED SPEECH
L dh	detalj therefore (*detail therefore*)	FULL
	bogalj themselves (*invalid themselves*)	NATIVE
	kasalj their (*cough their*)	PHONES
	ugalj there (*coal there*)	PAUSE
	temelj those (*foundation those*)	RECORDED SPEECH
ts ax	lanac append (*chain append*)	FULL
	punac approve (*father-in-law approve*)	NATIVE

Diphone	*Word Pair*	*Method*
	stranac attend (*foreigner attend*)	PHONES
	konac appoint (*thread appoint*)	PAUSE
	krivac apply (*guilty-person apply*)	RECORDED SPEECH

B.0.2 Bosnian - German

Diphone	*Word Pair*	*Method*
g p	prilog Partner (*contribution partner*)	FULL
	izlog Partner (*shop-window partner*)	NATIVE
	nalog Partei (*order party*)	PHONES
	vrtlog Parfüm (*whirl parfume*)	PAUSE
	zalog- Plastik (*pledge plastics*)	RECORDED SPEECH
J f	pucanj vierzig (*shot fourty*)	FULL
	stupanj Firma (*level company*)	NATIVE
	tutanj viertel (*roar(n.) quarter*)	PHONES
	svibanj vierzehn (*May fourteen*)	PAUSE
	pladanj Fehler (*tray mistake*)	RECORDED SPEECH
o E	krenuo Erde (*(he)-moved earth*)	FULL
	banuo Ernte (*(he)-burst-in harvest*)	NATIVE
	tonuo Erdnuss (*(he)-sank peanut*)	PHONES
	brinuo Erguss (*(he)-worried*)	PAUSE
	skinuo ertrank (*he-took-off(e.g. clothes) drowned*)	RECORDED SPEECH
tc pf	pomoć Pflanzen (*help plants*)	FULL
	ponoć Pflege (*midnight care*)	NATIVE
	sinoć Pflüge(*last-night ploughs*)	PHONES
	nemoć Pflichten (*weakness duty*)	PAUSE
	moguć Pfeife (*possible pipe*)	RECORDED SPEECH

Appendix C

Prompts

C.1 Prompts for the voice multiling_full_pause_multisyn

(fullpause_001 ”prilog Parade”)
(fullpause_002 ”izlog Partner”)
(fullpause_003 ”nalog Partei”)
(fullpause_004 ”vrtlog Parfuem”)
(fullpause_005 ”pucanj vierzig”)
(fullpause_006 ”stupanj Firma”)
(fullpause_007 ”tutanj vierzehn”)
(fullpause_008 ”svibanj Viertel”)
(fullpause_009 ”zalog Plastik”)
(fullpause_010 ”pladanj Fehler”)
(fullpause_011 ”skinuo ertrank”)
(fullpause_012 ”moguc Pfeife”)
(fullpause_013 ”krenuo Oede”)
(fullpause_014 ”banuo einsam”)
(fullpause_015 ”tonuo Paket”)
(fullpause_016 ”brinuo grosse”)
(fullpause_017 ”krenuo ”)
(fullpause_018 ”banuo ”)
(fullpause_019 ”tonuo ”)
(fullpause_020 ”brinuo ”)

(fullpause_021 ”pomoc anders”)
(fullpause_022 ”ponoc Zucker”)
(fullpause_023 ”sinoc neue”)
(fullpause_024 ”nemoc starke”)
(fullpause_025 ”pomoc ”)
(fullpause_026 ”ponoc ”)
(fullpause_027 ”sinoc ”)
(fullpause_028 ”nemoc ”)
(fullpause_029 ” Erde”)
(fullpause_030 ” Ernte”)
(fullpause_031 ” Erdnuss”)
(fullpause_032 ” Erguss”)
(fullpause_033 ” Pfluege”)
(fullpause_034 ” Pflege”)
(fullpause_035 ” Pflanzen”)
(fullpause_036 ” Pflichten”)
(fullpause_037 ”tabloa Erde”)
(fullpause_038 ”govor Ernte”)
(fullpause_039 ”daruj Erdnuss”)
(fullpause_040 ”istog Erguss ”)
(fullpause_041 ”bogalj Pfluege”)
(fullpause_042 ”svezanj Pflege”)
(fullpause_043 ”jasan Pflanzen”)
(fullpause_044 ”vodic Pflichten”)
(fullpause_045 ”talog party”)
(fullpause_046 ”razlog people”)
(fullpause_047 ”brlog pages ”)
(fullpause_048 ”zbog penny”)
(fullpause_049 ”toranj forty ”)
(fullpause_050 ”bubanj forward ”)
(fullpause_051 ”oganj formal”)
(fullpause_052 ”susanj forces”)
(fullpause_053 ”proslog programme”)
(fullpause_054 ”pladanj finding ”)
(fullpause_055 ”temelj those”)

(fullpause_056 ”krivac apply”)
(fullpause_057 ”detalj party”)
(fullpause_058 ”bogalj forward”)
(fullpause_059 ”kasalj advise”)
(fullpause_060 ”ugalj begin”)
(fullpause_061 ”lanac outbid”)
(fullpause_062 ”punac author”)
(fullpause_063 ”stranac metres”)
(fullpause_064 ”konac fewer”)
(fullpause_065 ”detalj ”)
(fullpause_066 ”bogalj ”)
(fullpause_067 ”kasalj ”)
(fullpause_068 ”ugalj ”)
(fullpause_069 ”lanac ”)
(fullpause_070 ”punac ”)
(fullpause_071 ”stranac ”)
(fullpause_072 ”konac ”)
(fullpause_073 ”saraf therefore”)
(fullpause_074 ”cekic themselves”)
(fullpause_075 ”krivio their”)
(fullpause_076 ”kraju there”)
(fullpause_077 ”oblik append”)
(fullpause_078 ”kastel approve”)
(fullpause_079 ”program attend”)
(fullpause_080 ”proces appoint”)
(fullpause_081 ” therefore”)
(fullpause_082 ” themselves”)
(fullpause_083 ” their”)
(fullpause_084 ” there”)
(fullpause_085 ” append”)
(fullpause_086 ” approve”)
(fullpause_087 ” attend”)
(fullpause_088 ” appoint”)

C.2 Prompts for the voice multiling_phones_multisyn

(phones_001 ”izlog”)
(phones_002 ”stupanj”)
(phones_003 ”banuo”)
(phones_004 ”ponoc”)
(phones_005 ”nalog”)
(phones_006 ”tutanj”)
(phones_007 ”tonuo”)
(phones_008 ”sinoc”)
(phones_009 ”razlog”)
(phones_010 ”bubanj”)
(phones_011 ”bogalj”)
(phones_012 ”punac”)
(phones_013 ”brlog”)
(phones_014 ”oganj”)
(phones_015 ”kasalj”)
(phones_016 ”krivac”)
(phones_017 ”Partner”)
(phones_018 ”vierzig”)
(phones_019 ”Ernte”)
(phones_020 ”Partei”)
(phones_021 ”Viertel”)
(phones_022 ”Erdnuss”)
(phones_023 ”Pfluege”)
(phones_024 ”party”)
(phones_025 ”forty”)
(phones_026 ”themselves”)
(phones_027 ”approve”)
(phones_028 ”penny”)
(phones_029 ”forces”)
(phones_030 ”their”)
(phones_031 ”apply”)
(phones_032 ”formal”)
(phones_033 ”susanj”)

(phones_034 ”stranac”)
(phones_035 ”attend”)

C.3 Prompts for the voice multiling_native_multisyn

(native_001 ”izlog pamet”)
(native_002 ”razlog partner”)
(native_003 ”stupanj firma”)
(native_004 ”bubanj fora”)
(native_005 ”bubanj foca”)
(native_006 ”Partner”)
(native_007 ”vierzig”)
(native_008 ”party”)
(native_009 ”forty”)
(native_010 ”banuo”)
(native_011 ”ponoc”)
(native_012 ”bogalj”)
(native_013 ”punac”)
(native_014 ”Ernte”)
(native_015 ”Pflege”)
(native_016 ”themselves”)
(native_017 ”approve”)
(native_018 ”bogalj demir”)
(native_019 ”punac apel”)
(native_020 ”banuo Ernes”)
(native_021 ”ponoc fleka”)

Bibliography

A., B. & Lenzo, K. (2000), Building voices in the festival speech synthesis system. http://festvox.org.

Badino, L., Barolo, C. & Quazza, S. (2004), Language independent phoneme mapping for foreign tts, *in* '5th ISCA Speech Synthesis Workshop', Carnegie Mellon University, Pittsburgh.

Bard, E., Robertson, D. & Sorace, A. (1996), 'Magnitude estimation of linguistic acceptability', *Language* **71**, 32–68.

Beutnagel, M. & Conkie, A. (1999), Interaction of units in a unit selection data base, *in* 'Proceedings of the 6th Conference on Speech Communication and Technology (Eurospeech 99)', Budapest, Hungary, pp. 1063–1066.

Black, A. & Lenzo, K. (2003), 'Optimal utterance selection for unit selection speech synthesis databases', *International Journal of Speech Technology* **6**(4), 357–363. Kluwer Academic Publishers.

Black, A., Taylor, P. & Caley, R. (2002), *The Festival Speech Synthesis System*, 1.4 edn.

Bozkurt, B., Ozturk, O. & Dutoit, T. (2003), Text design for tts speech corpus building using a modified greedy selection, *in* 'Proceedings of the 10th Conference on Speech Communication and Technology (Eurospeech 03)', Geneva, Switzerland.

Brabec, I., Hraste, M. & Živković (1952), *Gramatika Hrvatskoga ili Srpskoga Jezika*, Izdavačko poduzeće Školska knjiga.

Campbell, N. & Black, A. (1996), Prosody and the selection of source units for concatenative synthesis, *in* J. van Santen, R. W. Sproat, J. Olive & J. Hirschberg, eds, 'Progress in Speech Synthesis', Springer Verlag, Berlin, pp. 272–292.

Campbell, W. (2001), Talking foreign. concatenative speech synthesis and language barrier, *in* 'Proceedings of the 7th Conference on Speech Communication and Technology (Eurospeech 01)', Aalborg, Denmark.

Campbell, W. & Black, A. (1995), Optimising selection of units from speech databases for concatenative synthesis, *in* 'Proceedings of the 4th Conference on Speech Communication and Technology (Eurospeech 95)', Madrid, Spain.

Clark, R. A., Richmond, K. & King, S. (2004), Festival 2 – build your own general purpose unit selection speech synthesiser, *in* 'Proc. 5th ISCA workshop on speech synthesis'.

Clark, R. & King, S. (2003), Building a limited domain sythesiser with festival [online].

Corley, M., Corley, S., Keller, F., Konieczny, L. & Todirascu, A. (2004), Webexp experimental software. HCRC, University of Edinburgh, DFKI, University of Saarland.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F. & van der Vreken, O. (1996), The mbrola project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes, *in* 'Proceedings of ICSLP', Vol. 3, Philadelphia, pp. 1393–1396.

Eklund, R. & Lindström, A. (1996), Pronunciation in an internationalized society: a multi-dimensional problem considered, *in* 'FONETIK 96, Swedish Phonetics Conference', Nässlingen, Sweden, pp. 123–126.

Eklund, R. & Lindström, A. (1998), How to handle "foreign" sounds in swedish text-to-speech conversion: Approaching the "xenophone" problem, *in* 'Proceedings of the 5th International Conference on Spoken Language Processing', Sydney, Australia, pp. 2831–2835.

Eklund, R. & Lindström, A. (1999), Xenophones revisited: linguistic and other underlying factors affecting the pronunciation of foreign items in swedish, *in* 'Proceedings of ICPhS 99', Vol. 3, San Francisco, California, pp. 2227–2230.

Eklund, R. & Lindström, A. (2000), How foreign are "foreign" speech sounds? implications for speech recognition and speech synthesis, *in* 'Proceedings of the RTO

Meeting, Multi-Lingual Interoperability in Speech Technology', Hull (Quebec), Canada, pp. 15–19.

François, H. & Boëffard, O. (2001), Design of an optimal continuous speech database for text-to-speech synthesis considered as a set covering problem, *in* 'Proceedings of the 7th Conference on Speech Communication and Technology (Eurospeech 01)', Aalborg, Denmark, pp. 829–832.

Holmes, J. & Holmes, W., eds (2001), *Speech Synthesis and Recognition*, Taylor and Francis.

Hunt, A. & Black, A. (1996), Unit selection in a concatenative speech synthesis system using a large speech database, *in* 'Proceedings of ICASSP 96', Vol. 1, Atlanta, Georgia, pp. 373–376.

Hunt, M., Zwierynski, D. & Carr, R. (1989), Issues in high quality lpc analysis and synthesis, *in* 'Proceedings of the 1st Conference on Speech Communication and Technology (Eurospeech 89)', Vol. 2, Paris, France, pp. 348–351.

Ivić, P. (1958), *Die serbokroatischen Dialekte: Ihre Struktur und Entwicklung. Erster Band: Allgemeines und die štokavische Dialektgruppe*, Mouton, The Hague.

Kishore, S. & Black, A. (2003), Unit size in unit selection speech synthesis, *in* 'Eurospeech 03', Geneva, Switzerland.

Klatt, D. (1975), 'Vowel lengthening is syntactically determined in a connected discourse', *Journal of Phonetics* **3**, 129–140.

Lehiste, I., Olive, J. & Streeter, L. (1976), 'The role of duration in disambiguating syntactically ambiguous sentences', *Journal of Acoustical Society of America* (60), 1199–1202.

Makashay, M., Wightman, C., Syrdal, A. & Conkie, A. (2000), Perceptual evaluation of automatic segmentation in text-to-speech synthesis, *in* 'IC-SLP2000', Beijing, China.

Möbius, B., Sproat, R., van Santen, J. & Olive, J. (1997), The bell labs german text-to-speech system: An overview, *in* 'Proceedings of the 5th Conference on Speech Communication and Technology (Eurospeech 97)', Rhodes, Greece, pp. 2443–2446.

Moulines, E. & Charpentier, F. (1990), 'Pitch-synchronous waveform processing

techniques for text-to-speech synthesis using diphones', *Speech Communication* **9**(5), 453–467.

Olive, J., van Santen, J., Mobius, B. & Shih, C. (1998), *Multilingual Text-to-Speech Synthesis, The Bell Labs Approach*, Kluwer Academic Publishers, chapter 7, pp. 191–228.

Remijsen, B. & van Heuven, V. (2004), 'Word prosody of papiamentu', *Phonetics* . under review.

Saikachi, Y. (2003), Building a unit selection voice for festival, Master's thesis, University of Edinburgh.

Sluijter, A. & van Heuven, V. (1996), 'Spectral balance as an acoustic correlate of linguistic stress', *Journal of the Acoustical Society of America* **100**(4), 2471–2485.

Sorace, A. (2003), Magnitude estimation of linguistic acceptability: applications to research on developing grammars. Unpublished presentation at the University of Utrecht.

Stöber, K., Poerterle, T., Wagner, P. & W., H. (1999), Synthesis by word concatenation, *in* 'Proceedings of the 6th Conference on Speech Communication and Technology (Eurospeech 99)'.

Traber, C., Huber, K., Nedir, K., Pfister, B., Keller, E. & Zellner, B. (1999), From multilingual to polyglot speech synthesis, *in* 'Proceedings of the 6th Conference on Speech Communication and Technology (Eurospeech 99)', Budapest, Hungary, pp. 835–838.

van Santen, J. (1997), Combinatorial issues in text-to-speech synthesis, *in* 'Eurospeech 97', Vol. 2, Rhodes, Greece.

van Santen, J. & Buchsbaum, A. (1997), Methods for optimal text selection, *in* 'Eurospeech 97', Rhodes, Greece.

Yi, J. & Glass, J. (1998), Natural-sounding speech synthesis using variable-length units, *in* 'Proc. ICSLP-98', Vol. 4, Sydney, Australia, pp. 1167–1170.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. & P., W. (2002), *The HTK Book (for HTK version 3.2)*, Cambridge Universtiy Engeneering Department.