UDK 004.934:004.5=163.42

RAČUNALNI SUSTAV ZA TVORBU HRVATSKOGA GOVORA
TEXT-TO-SPEECH SYNTHESIS: A PROTOTYPE SYSTEM FOR CROATIAN LANGUAGE

Miran POBAR – Sanda MARTINČIĆ-IPŠIĆ – Ivo IPŠIĆ

Sažetak: *U radu je prikazan sustav koji omogućuje umjetnu tvorbu hrvatskoga govora prema proizvoljnom ulaznom tekstu. Ulazni tekst, koji mora biti u normaliziranom obliku, sustav pretvara u niz fonema (pretvorba grafem-fonem), a zatim stvara zvučni zapis na temelju fonetskoga niza. Korišteni postupak sinteze temelji se na ulančavanju manjih akustičkih jedinica govora – difona metodom TD-PSOLA. Za potrebe sustava izrađena je i baza difona za hrvatski govor. Predložen je automatski postupak odabira difona iz govornoga korpusa.*
Kvaliteta ostvarenoga postupka ispitana je provođenjem ankete među ispitanicima. Ispitanici su dali subjektivnu ocjenu kvalitete dobivenoga govora, a time je provjerena i njegova razumljivost.

Ključne riječi:
- odabir difona
- baza difona za hrvatski jezik
- ulančavanje difona
- umjetna tvorba govora
- procjena kvalitete govora

Abstract: *This paper presents the development of a Croatian text-to-speech system capable of synthesizing speech from arbitrary text. Input text in normalized form is first transcribed into a phonetic string (grapheme-to-phoneme conversion) and then processed by a TD-PSOLA based synthesizer. A procedure for automatic selection of diphones from a spoken corpus is proposed. A Croatian language diphone database was built for the system. Subjective quality evaluations of the resulting speech were performed, as well as tests for intelligibility.*

Keywords:
- diphone selection
- Croatian diphone database
- diphone concatenation
- text-to-speech systems
- speech quality evaluation

1. UVOD

Sustavi za sintezu govora vrše pretvorbu ulaznoga teksta (niza riječi odnosno rečenica) u govor. Sinteza govora sastavni je dio govornih tehnologija, koje omogućuju govornu komunikaciju čovjeka i strojeva. Jedna od primjena sinteze govora su sustavi koji omogućuju slijepim i slabovidnim osobama pristup sadržajima u elektroničkom obliku. Primjenjuju se i u sustavima u kojima oči i ruke trebaju biti slobodne, primjerice kod sustava za navigaciju u automobilu. Telefonski sustavi koji opslužuju velik broj korisnika informacijama poput voznih redova mogu se automatizirati korištenjem takvih sustava. U sprezi sa sustavima za raspoznavanje govora u nekim slučajevima mogu biti prikladniji kao korisničko sučelje od uobičajeno korištenih tipkovnica i ekrana, osobito na malenim uređajima poput mobilnih telefona i ručnih računala.
Postupaka tvorbe zvučnih (govornih) signala na temelju

1. INTRODUCTION

Text-to-speech (TTS) systems transform input text (sequence of words or sentences) into speech. Speech synthesis is a part of speech technologies, which enable spoken man-machine communication. One of the applications of TTS systems allows blind and visually impaired persons access to written electronic content. Such users can easily obtain web pages, e-mail messages and generally electronic documents in spoken form. TTS systems are also used in eyes- and hands-free applications, such as in-car navigation systems. Systems for providing information over the telephone, for example the train timetable, can be automated using TTS. TTS systems used together with speech recognition may in some cases be better suited as a user interface than standard keyboards and screens, especially on small devices like cell phones and handheld computers.
There are several methods for synthesizing a speech

određenoga fonetskog niza i načina izgovora ima nekoliko, a međusobno se razlikuju prema razumljivosti, prirodnosti, fleksibilnosti, zahtjevima za strojnom podrškom i vremenu koje je potrebno uložiti za njihovu implementaciju. Odabir metode ovisi o mogućoj primjeni sustava i očekivanoj kvaliteti govora.
Rad je organiziran na sljedeći način: drugo poglavlje prikazuje građu sustava za umjetnu tvorbu govora, u trećem poglavlju opisane su metode umjetne tvorbe govora, a u četvrtom je pobliže prikazana metoda tvorbe ulančavanjem segmenata prirodnoga govora. Peto poglavlje donosi opis izrade baze difona za hrvatski govor, a šesto opis izrade postupka sinteze govora u Matlabu. Na kraju slijede ocjena evaluacije kvalitete dobivenoga govora te zaključak.

2. SUSTAVI ZA UMJETNU TVORBU GOVORA

Višejezični sustav Festival [1][2] cjelovit je sustav za pretvorbu teksta u govor (TTS), a za sada podržava engleski, španjolski i velški jezik. Građen je modularno, što omogućuje korištenje različitih postupaka sinteze. Jedan od takvih postupaka je MBROLA [3] koji se temelji na ulančavanju snimljenih govornih segmenata – difona. Za MBROLA sintetizator razvijena je baza difona i za hrvatski jezik [4]. Postupak sinteze temeljen na skrivenim Markovljevim modelima govora [5][6] također omogućuje razvoj sustava za različite jezike uz odgovarajući govorni korpus. Za srpski i slovenski jezik moguće je naći sustave za umjetnu tvorbu govora koji za sintezu koriste TD-PSOLA [7] algoritam [8][9].

2.1. Struktura sustava za umjetnu tvorbu govora

Na slici 1 shematski je prikazana općenita struktura sustava za umjetnu tvorbu govora i pripadnih modula, prema [10].

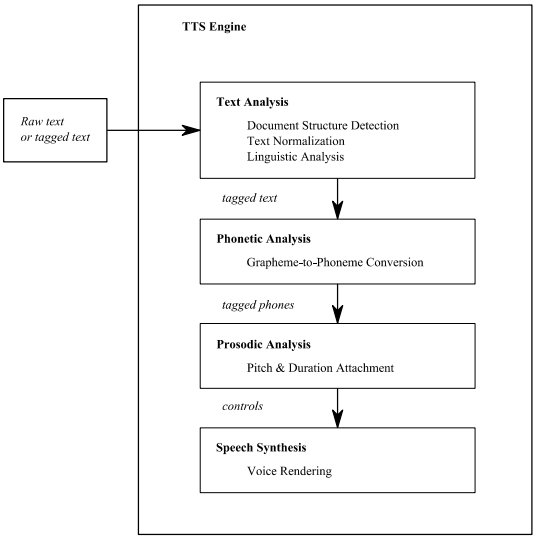
signal from a given phonetic string and prosody, each with specific characteristics regarding intelligibility, naturalness, flexibility, system requirements and time required for implementation. Choice of the used method will depend on the system application and the expected voice quality.
This paper is organized as follows: the second chapter describes the architecture of the TTS system; the third chapter describes methods for speech synthesis while the fourth describes diphone concatenation synthesis. The fifth chapter describes the diphone database for Croatian speech synthesis. In chapter six we describe the speech synthesis procedure developed in Matlab. In the end we present the evaluation results and give some conclusions.

2. TEXT-TO-SPEECH SYSTEMS

The multi-lingual system Festival [1][2] is a complete TTS system with support for English, Spanish and Welsh languages. It has a modular structure that facilitates switching between different speech synthesis methods. One of such methods is MBROLA [3], based on concatenation of recorded speech segments-diphones. A Croatian language diphone database was also developed for the MBROLA synthesizer [4]. Hidden Markov model based speech synthesis [5][6] can be used for different languages if a speech corpus for a selected language exists. For Serbian and Slovene languages TD-PSOLA [7] based TTS systems can be found in [8][9].

2.1 TTS System Structure

Figure 1 shows the general structure of a modular text-to-speech system, as described in [10].



Slika 1. Shematski prikaz sustava za sintezu govora.
Figure 1. Block diagram of a TTS system.

Modul za analizu teksta ima zadatak pročistiti ulazni tekst od svih elemenata koji nisu jednoznačno izgovorljivi (normalizacija). Brojeve i ostale simbole treba raspisati u riječi, a kratice proširiti u puni oblik. Primjerice tekst «2. sv. rat trajao je» treba prepisati u «Drugi svjetski rat trajao je». Taj se postupak najčešće vrši opsežnim skupom ručno zadanih pravila koja ipak ne mogu pokriti sve slučajeve zbog golema broja mogućih formata koji se u tekstu mogu pojaviti, pogrešaka u tekstu koje mogu «zavarati» pravila, a u nekim primjerima višesmislenosti ni čovjek ne može jednoznačno odrediti kako nešto treba biti pročitano.
Ovaj modul ima i zadatak prepoznavanja strukture dokumenta koja utječe na način izgovora (prozodiju), a osobito na stanke u govoru. Ulazni tekst može sadržavati i neke oznake ili upute sintetizatoru, kao što su željeni spol govornika, brzina izgovora i visina tona.
Zadatak modula za fonetsku analizu je pretvorba normaliziranog teksta u odgovarajući fonetski zapis. Za hrvatski jezik ta je pretvorba jednostavna, a vrši se pravilima za zamjenu grafema odgovarajućim fonemskim parom. Za iznimke i strane riječi koristi se fonetski rječnik.
Cilj je modula za prozodijsku analizu stvoriti parametre za sintezu koji će dati najprirodniji izgovor zadane rečenice ili iskaza. Pod pojmom *prozodija* podrazumijevaju se fonetski učinci poput glasnoće, visine glasa, stanki u govoru, brzine i ritma. Visina odnosno ton glasa opisuje se osnovnom frekvencijom *F0* izraženom u Hz koja predstavlja osnovnu frekvenciju titraja glasnica. Uz osnovnu frekvenciju za svaki se izgovoreni fon bilježi trajanje i eventualno glasnoća.

3. METODE UMJETNE TVORBE GOVORA

3.1. Artikulacijski model tvorbe

Artikulacijski model tvorbe govora temelji se na matematičkom modelu kojim se opisuju oblik govornoga trakta, mehanička kretanja artikulatora i kretanje zračne struje govornim traktom. Parametri modela mogu se dobiti iz rendgenskih snimki ili snimki magnetskom rezonancijom prema stvarnome ljudskom govornom traktu prilikom izgovora glasova. Kvaliteta govora dobivena tom metodom nije u dosadašnjim realizacijama na razini one dobivene metodom formanata ili ulančavanjem segmenata prirodnoga govora [10].

3.2. Tvorba metodom formanata

Glasove je moguće sintetizirati propuštanjem odgovarajućih pobudnih signala kroz linearni filtar podešen prema rezonantnim frekvencijama ljudskoga govornog trakta za željeni glas. Pobuda za zvučne glasove je niz impulsa, za bezvučne bijeli šum, a za zvučne frikative istodobno niz impulsa i šum.
Ta je metoda tvorbe govora vrlo fleksibilna što se tiče

A text analysis module processes the input text, transcribing numbers, symbols, acronyms and other elements that cannot be unambiguously pronounced into words. This is called text normalization. For example, "2. sv. rat trajao je" is transcribed as "drugi svjetski rat trajao je". This process is commonly done using an extensive set of manual rules that still cannot cover all the possible formats that may appear in a text source. Additionally typing errors can influence the process of normalization. In some cases it is even difficult for human speakers to determine the correct pronunciation.
This module also detects the document structure which has an important effect on prosody, especially on pauses. Input text may also contain tags or commands to the synthesizer, such as speaker's sex, reading speed and pitch range.
The phonetic analysis module transforms the normalized text into the corresponding phonetic string. For Croatian this is a simple process and can be done with rules that map the graphemes to their phonetic pairs. Exceptions and foreign words are handled by a phonetic dictionary.
The goal of the prosodic analysis module is to create parameters for synthesis that will give the most natural output for a given sentence. These parameters include volume, pitch, pauses, speed and rhythm. Pitch is represented with fundamental frequency of glottal pulses *F0* in Hertz. Besides the fundamental frequency, phoneme duration and sometimes the volume is estimated.

3. SPEECH SYNTHESIS PROCEDURES

3.1 Articulatory Speech Synthesis

Articulatory speech synthesis uses a mathematical model that describes the form of the human vocal tract, mechanical movements of the articulators and the airflow in the vocal tract. The model parameters are obtained from X-rays or magnetic resonance imaging (MRI) of the actual vocal tract during speech production. The quality of synthesized speech using this method has so far been lower than the one produced by formant model synthesis or concatenative synthesis [10].

3.2. Formant Model

Sounds can be synthesized by passing certain waveforms through a linear filter. An impulse train is used for voiced sounds, white noise for unvoiced sounds and a combination of the two for voiced fricatives. The linear filter is set to formant frequencies of the human vocal tract corresponding to the desired sound.
This method is very flexible in terms of output voice

karakteristika izlaznoga glasa i omogućuje dobro praćenje zadanih prozodijskih karakteristika. Kvaliteta ostvarenoga govora najčešće udovoljava kriteriju razumljivosti, ali zvuči neprirodno [11].

3.3. Tvorba na osnovi statističkoga modela govora (HMM sinteza)

Umjesto ručno zadanih pravila kojima se određuju vektori parametara glasa kod sinteze govora metodom formanata, može se koristiti metoda koja se temelji na skrivenim Markovljevim modelima (HMM) [12].
Trifoni, odnosno glasovi ovisni o artikulacijskom kontekstu, modeliraju se skrivenim Markovljevim modelima. Sustav najprije u fazi učenja analizira snimljeni govor i odgovarajuću fonetsku transkripciju na temelju kojih se generiraju vektori karakteristika mel-kepstra, osnovne frekvencije F0 i njihovih dinamičkih osobina. Oni se koriste za učenje HMM-a koje rezultira kontekstno ovisnim akustičkim modelom. Na temelju dobivenoga akustičkog modela i ulaznoga teksta stvaraju se parametri za sintezu govora koja se vrši pomoću izvor-filtar modela. Prednost je toga postupka što se na temelju prirodnoga govora bilježe svojstva karakteristična za kontekst u kojem se glasovi izgovaraju.

4. TVORBA GOVORA ULANČAVANJEM

Kod tvorbe govora ulančavanjem govorni se isječak sintetizira jednostavnim reproduciranjem prethodno snimljenog segmenta prirodnoga govora. Izgovor željenog iskaza dobiva se spajanjem potrebnoga broja takvih segmenata koji odgovaraju zadanome fonetskom nizu. Ako su korišteni isječci izvađeni iz različitih okruženja, na mjestima spoja može doći do izobličenja. Takvi isječci mogu imati i posve različitu intonaciju koja se u prirodnom govoru ne može pojaviti.. Njihovim spajanjem dobiveni sintetizirani govor zvučat će neprirodno, iako svaki zasebni isječak zvuči posve prirodno. Ako pak za zadanu frazu u bazi postoje isječci koji dobro pristaju jedan uz drugi, umjetno stvoreni govor može zvučati gotovo jednako dobro kao prirodni.

4.1. Baza akustičnih jedinica

4.1.1. Akustične jedinice – difoni

Akustične jedinice na osnovi kojih se može sintetizirati govor ulančavanjem su fonemi, podfonetske jedinice, difoni, trifoni, slogovi, riječi i fraze. Odabirom duljih elemenata poput riječi i fraza, smanjuje se broj mjesta spajanja i time broj mjesta na kojima mogu nastati izobličenja. U tom je slučaju moguća tvorba samo iz ograničene domene onih isječaka koji postoje u bazi. Za tvorbu govora prema proizvoljnom tekstu prikladniji su kraći elementi poput *difona*. Difon O-B u riječi *oblak* jedinica je koja se proteže od sredine trajanja glasa O do sredine trajanja glasa B. U nekom jeziku koji ima *N*

characteristics and the synthesized speech can closely follow the desired prosody. The resulting speech is usually very intelligible, but can sound unnatural [11].

3.3 Synthesis Using Statistical Speech Models (HMM)

Instead of using manually configured rules for determining the speech signal features, a hidden Markov models method can be used [12].
Triphones or context dependent phones are modeled using hidden Markov models. In the training phase, the system analyses the recorded speech and corresponding phonetic transcription, generating feature vectors for mel-cepstrum, fundamental frequency and their dynamic features. These vectors are used in training HMMs and obtaining the context-dependent acoustic model.
Using the derived model and input text, speech synthesis parameters are generated and passed on to a source-filter model synthesizer. The advantage of this method is that the context-dependent properties of the voices are derived from natural speech.

4. CONCATENATIVE SPEECH SYNTHESIS

In concatenative speech synthesis a speech segment is synthesized by simply reproducing segments of previously recorded natural speech. A desired utterance is synthesized by reproducing a sequence of such segments that correspond to the given phonetic string. If segments are extracted from a different context, distortion can occur on concatenation points. These segments can also have completely different prosodic characteristics that never come together in natural speech. Speech synthesized using such segments can sound unnatural even though each segment is completely natural. If there are well-fitting segments in the database for synthesizing the desired utterance, synthetic speech can sound almost as good as natural speech.

4.1. Acoustic Units Database

4.1.1 Acoustic Units – Diphones

Acoustic units used in concatenative speech synthesis are phonemes, subphonetic units, diphones, triphones, syllables, words and phrases. The number of needed concatenation points and possible places where distortion occurs is reduced when longer units such as words and phrases are used. In that case it is only possible to synthesize speech from a limited domain, determined by the speech segment database inventory. Shorter segments such as diphones are more suitable for synthesizing arbitrary speech. The diphone O-B in the word *oblak* is a unit that spans from the middle of the voice o to the

fonema, difona ima teorijski N^2, no u govoru se ne pojavljuju svi pa ih je u bazi potreban manji broj. U hrvatskom jeziku fonema ima N = 30 pa je za tvorbu proizvoljne riječi potrebno manje od 900 difona.
Izgovor dobiven ulančavanjem segmenata ne može pratiti prozodiju zadanu u prethodnom modulu osim ako za sve difone nisu pohranjene instance (izgovori difona) za svaki naglasak. Zamjena silaznoga naglaska uzlaznim ili obrnuto dat će neprirodan izgovor, a može doći i do promjene značenja riječi. Potreba za pohranjivanjem velikog broja instanci istog difona može se izbjeći korištenjem neke od metoda za modifikaciju prozodije, kao što je PSOLA [7].

middle of the voice b. In a language that has N phonemes, there are N^2 possible diphones but not all of them appear in speech so a smaller number of diphones is needed in the database. There are N=30 phonemes in the Croatian language so we need less than 900 diphones to form an arbitrary word. Unless we keep an instance for every accent of each diphone in the database, the prosody of the synthesized speech cannot follow the one set by the prosodic analysis module. Substitution of rising with falling accent or vice versa will sound unnatural and may even change the meaning of the word. The need to store many instances of diphones can be avoided by using some of the prosody modification methods such as PSOLA [7].

4.1.2. Izrada baze akustičkih jedinica – difona

Izgradnja inventara akustičkih jedinica počinje snimanjem govora istoga govornika u što stalnijim uvjetima. Potrebno je prikupiti dovoljno snimki da se dobije što bolja pokrivenost difona. Druga je mogućnost da se izgovaraju rečenice u koje su umetnute riječi bez značenja – logatomi, smišljeni tako da sadrže željene difone. Logatomi su sastavljeni od četiriju fonema, iz kojih se «izrezuje» središnji difon, npr. iz logatoma *abda* dobije se difon *bd* [4].
Za prikupljene govorne zapise potrebno je generirati fonetski prijepis s točnim vremenima početka i kraja svakoga fonema. Na temelju dobivenih vremenskih odsječaka «izrezuju» se difoni iz izvornoga zvučnog zapisa.

4.1.2 Acoustic Unit Database Construction

Building the inventory of the acoustic units begins with the recording of speech. All recorded material must be from the same speaker and recorded in similar conditions. It is necessary to collect enough recordings in order to get good coverage of all diphones. Another option is to record sentences that contain meaningless words – logatoms, specifically made to contain the desired diphones. Logatoms are made of four phonemes from which the middle diphone is extracted, e.g. from the logatom *abda* we extract the diphone *bd* [4]. Phonetic transcription with precise timing of the beginning and the ending of each voice must then be done manually for the collected recordings. Diphones can be extracted from speech recordings using those timings.

4.2. Spajanje segmenata postupcima OLA i PSOLA

Postupci OLA («overlap and add», *preklopi i zbroji*) i TD-PSOLA(«time-domain pitch synchronous overlap and add») [7] koriste se za modifikaciju prozodije i spajanje zvučnih segmenata kod tvorbe govora ulančavanjem. Metoda OLA omogućuje skraćivanje trajanja govora bez promjene visine tona. Za željeni faktor skaliranja f, ulazni signal $\mathbf{x}[n]$ množi se Hannovim vremenskim prozorima širine $2N$ uzoraka i razmaknutih fN uzoraka. Hannov prozor širine N uzoraka definiran je izrazom:

4.2. Segment Concatenation Using OLA and PSOLA

OLA (overlap-and-add) and PSOLA (pitch synchronous overlap-and-add) are used for prosody modification and joining of segments in concatenative speech synthesis. The OLA method allows shortening the duration of speech without altering the pitch. For the desired scaling factor f, the input signal $\mathbf{x}[n]$ is multiplied by Hanning windows of length $2N$ and spaced fN samples. A Hanning window of length N samples is defined by the expression:

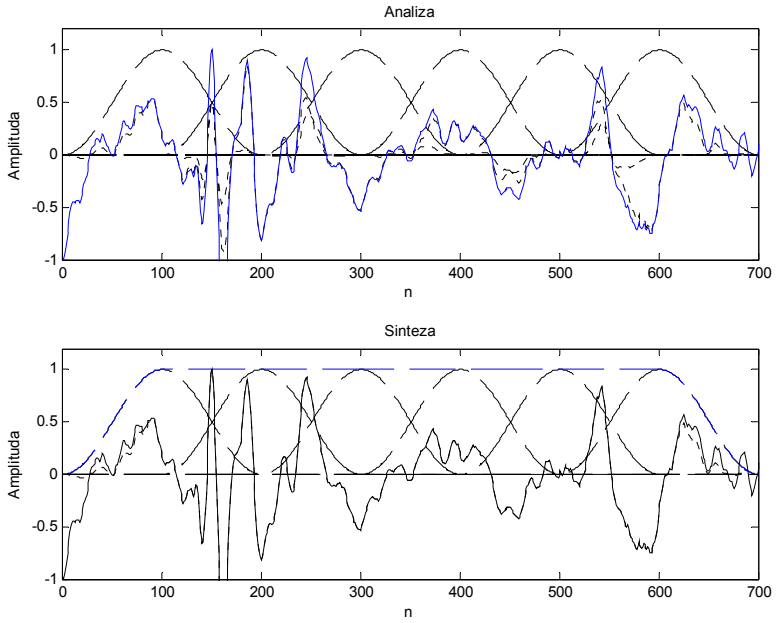
$$h[n] = \begin{cases} 0.5 - 0.5\cos\left(\dfrac{2\pi n}{N}\right) & 0 \le n < N \\ 0 & \text{inace/otherwise} \end{cases} \quad (1)$$

Središta se dobivenih signala pri sintezi razmiču na N uzoraka i njihove se vrijednosti zbrajaju. Za slučaj faktora skaliranja f=1 i korištenja Hannova prozora pri sintezi se dobiva neizmijenjen ulazni signal. Taj se slučaj može objasniti slikom 2. Gornja slika prikazuje 6 analitičkih prozora (isprekidana linija) kojima se množi ulazni signal (deblja linija). Prikazani su i rezultirajući kratkotrajni signali (crtkana linija). Donja slika prikazuje prozore

In synthesis the resulting signals are overlapped with their centers spaced N samples apart and summed together. If the scaling factor f is f=1 and the Hanning window is used, the synthesized signal is identical to the input signal, as shown in Figure 2. The upper graph shows the input signal (thicker line) that is multiplied by 6 analysis windows (dashed line). The dotted line represents the resulting short-time signals. The lower graph shows the

sinteze (isprekidana linija) i njihov zbroj (deblja isprekidana linija). Izvorni je signal prikazan punom crtom, a rekonstruirani isprekidanom.

synthesis windows (dashed line) and their sum (thicker dashed line). The reconstructed signal is shown with a dotted line and the source signal with the continuous line.



Slika 2. Rekonstrukcija signala metodom OLA uz f=1.
Figure 2. Signal reconstruction using OLA method with f=1.

Pri faktorima skaliranja različitima od 1 u sintetiziranom signalu može doći do neželjenoga gubitka periodičnosti signala zbog stalnog razmještaja analitičkih prozora, neusklađenog s osnovnom frekvencijom *F0*.
Poboljšanje te metode je PSOLA gdje su širina i razmak analitičkih prozora usklađeni s periodom osnovne frekvencije *F0* ulaznoga signala. Ta metoda osim manjih izobličenja pri spajanju difona omogućuje i promjenu tona odnosno visine glasa. Primjena postupka PSOLA zahtijeva poznavanje vremenske ovisnosti *F0* u signalu, što nije bilo potrebno kod metode OLA.
Ulazni diskretni zvučni signal $\mathbf{x}[n]$ može se predstaviti zbrojem slijeda kratkih signala $x_i[n]$ dobivenih množenjem signala analitičkim prozorima $h_i[n]$:

Unwanted loss of periodicity may occur in the synthesized signal with a scaling factor different than 1, caused by constant spacing of analysis windows unsynchronized with an input signal fundamental frequency *F0*.
An improvement of the OLA method is PSOLA, where the analysis window width and spacing are synchronized with input signal pitch periods *F0*. Besides having less distortion when concatenating diphones than with OLA this method also allows changing of voice pitch. *F0* contour of the input signal must be known to apply this method, which is not required for OLA. The input discrete signal $\mathbf{x}[n]$ can be represented as a sum of short signal $x_i[n]$ sequences obtained by multiplying the input signal with the analysis window $h_i[n]$.

$$x_i[n] = h_i[t_a[i] - n]x[n] \tag{2}$$

Prozori su centrirani oko slijeda *vremenskih oznaka* $t_a[i]$, koje su na zvučnim dijelovima govora raspodijeljene sinkrono s periodima osnovne frekvencije glasa, a na bezvučnim dijelovima na jednake razmake, manje od 10 ms. Za zvučne dijelove razlika je između susjednih vremenskih oznaka $P_a[i] = t_a[i] - t_a[i-1]$ jednaka periodu P osnovne frekvencije signala u trenutku $t_a[i]$. Prozori $h_i[n]$ su Hannova tipa i uvijek su širi od lokalnog perioda P, obično proporcionalno s faktorom 2.

The windows are centered around a series of pitch marks $t_a[i]$, synchronously distributed with pitch periods of signal for voiced parts and evenly distributed by periods shorter than 10 ms on unvoiced parts. The difference between neighboring pitch marks for voiced parts $P_a[i] = t_a[i] - t_a[i-1]$ is equal to the pitch period of the signal in $t_a[i]$. The windows $h_i[n]$ are of Hanning type, always wider than the local pitch period P, usually by a factor of 2. The sequence of frames $x_i[n]$ is transformed

Niz dobivenih okvira $x_i[n]$ pretvara se u niz izlaznih okvira $y_j[n]$ sinkroniziran na nove vremenske oznake $t_s[j]$. Postupak promjene visine tona i trajanja opisan je u [7] i [10]. Zbrajanjem sekvence izlaznih okvira dobiva se izlazni signal $y[n]$:

into a series of output frames $y_j[n]$ synchronized to the new pitch marks $t_s[j]$. A method for changing the pitch and duration is described in [7], [10].By summing the output frames, the obtained sequence is the output signal $y[n]$:

$$y[n] = \sum_{j=-\infty}^{\infty} y_j[n - t_s[j]]. \tag{3}$$

4.2.1. Detekcija vremenskih oznaka

U dosadašnjem razmatranju pretpostavilo se da su vremenske oznake $t_a[i]$ poznate, dok ih u praksi treba odrediti na temelju samog govornog signala. Računske metode detekcije najčešće se temelje na praćenju osnovne frekvencije *F0*. Perioda $P_a(t)$ *F0* je jednoznačno određena vremenskim oznakama, prema izrazu $P_a(t) = t_a[i+1] - t_a[i]$, no obratno ne vrijedi zbog neodređenoga vremenskog ishodišta. Vremenske oznake stoga su određene samo približno.
Često korištena metoda za procjenu *F0* zasniva se na traženju najveće vrijednosti funkcije autokorelacije [10].
Unaprijeđene metode koje uzimaju u obzir činjenicu da se u prirodnom govoru *F0* ne mijenja naglo i mogu izgladiti dobivenu krivulju *F0* opisani su u [13],[10].

4.2.1 Pitch Marks Detection

So far, we have assumed that the pitch marks $t_a[i]$ are known, while in practice they must be determined from the input signal itself. Numerical pitch detection methods are usually based on pitch tracking. The pitch period $P_a(t)$ is exactly determined by the expression $P_a(t) = t_a[i+1] - t_a[i]$ while the pitch marks cannot be determined from the pitch period because of undetermined time origin. Pitch marks are thus only approximately determined. A frequently used pitch detection method is based on finding the maximum of the autocorrelation function [10]. More advanced methods take into account the fact that pitch doesn't change abruptly in natural speech and can smooth the obtained pitch contour as described in [13],[10].

5. PRIPREMA BAZE DIFONA ZA HRVATSKI JEZIK

Za potrebe sustava izrađena je baza difona za hrvatski jezik. Prije izrade baze definiran je skup fonema koje će sustav razlikovati. Osim 30 standardnih fonema za hrvatski jezik dodatno se razlikuju naglašene inačice samoglasnika (*a:, e:*, *i:, o:* i *u:*), pojava glasa *r* u funkciji samoglasnika, tišina kao poseban fonem ili sveukupno 37 fonema opisanih u [14].
Na temelju toga skupa teorijski je moguće razlikovati 37*37=1369 difona koji odgovaraju svim kombinacijama parova zadanih fonema.
Za izradu baze upotrebljeni su unaprijed prikupljeni govorni zapisi jednoga govornika iz govornoga korpusa hrvatskoga govora koji je sastavljen pri Odsjeku za informatiku Filozofskoga fakulteta u Rijeci [14]. Korišteni su isječci snimljenih radijskih emisija informativnoga sadržaja – vijesti i vremenske prognoze. Format zapisa je jednokanalni *.wav* frekvencije uzorkovanja 16 kHz. Ukupni broj zapisa je 2331, veličina 231 MB, a trajanje 2 sata i 26 min. Zvučne datoteke prethodno su bile obrađene postupkom automatskoga prepoznavanja govora, čime su generirane *.lab* datoteke s oznakama izgovorenih fona i automatski određenim trenucima početka i kraja svakoga fonema izraženima u milisekundama. Uz zvučne datoteke i oznake dostupni su bili i prijepisi govora te fonetski rječnik s 9922 unosa.

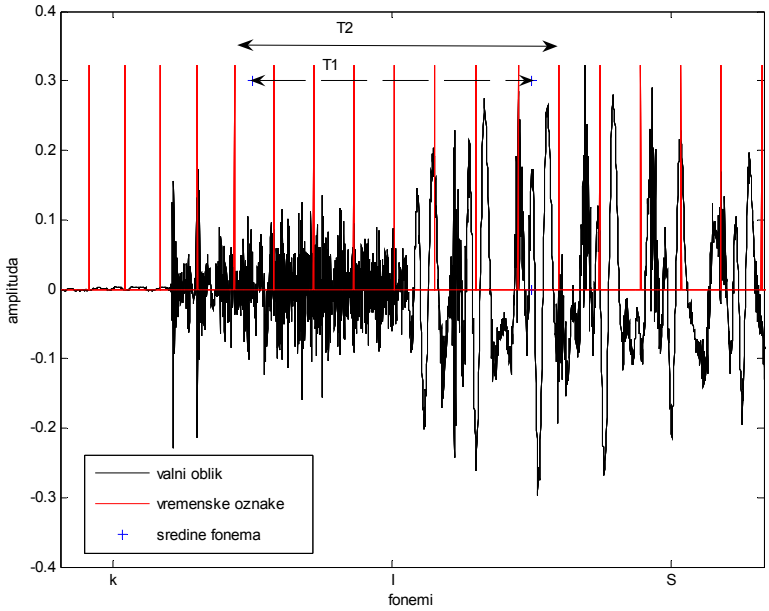
5. CROATIAN DIPHONE DATABASE PREPARATION

A Croatian language diphone database was built for the system. In addition to the 30 standard phonemes in the Croatian language, accented variants of the vowels (*a:, e:*, *i:, o:* and *u:*), the sound r as a vowel and silence as a distinguished phoneme, or a total of 37 phonemes, described in [14].

It is possible to differentiate 37*37=1369 diphones based on this set, corresponding to all possible phoneme pairs.
To build the diphone database we used previously collected speech recordings from one speaker from the speech corpus of Croatian speech [14]. We used samples from radio broadcasts of news and weather reports. Audio format of recordings is single channel .wav with sampling rate fs=16kHz. The total number of recordings is 2331, the size is 231 MB and the duration is 2h 26 min. The sound files were previously processed using an automated speech recognition system that generated phone labels and automatically determined time segments of each phone in milliseconds. Apart from sound and label files, there were available textual transcriptions of the recordings as well as a phonetic dictionary with 9922 entries. To simplify processing, symbols used in labeling, textual and dictionary files were unified. We used symbols from the ASCII set. Label files were analyzed

Radi jednostavnosti obrade ujednačeni su znakovi korišteni u datotekama s oznakama, prijepisima i rječniku. Korišteni su znakovi koji pripadaju ASCII skupu. Pomoću za tu svrhu izrađene funkcije *Matlab*, analizirane su sve .lab datoteke kako bi se utvrdila učestalost pojave pojedinih difona u snimljenom govoru.
S barem jednim pojavljivanjem prisutno je 1057 od teorijski mogućih 1396 difona.
U sljedećem koraku izvorni govorni segmenti podijeljeni su u difone. Difoni su pohranjivani kao .wav datoteke, imenovane automatski u obliku "1.fonem-2.fonem-redni_broj-riječ_iz_koje_je_izrezan.wav", npr. "a-d-11-nadajući.wav". Kako bi na spojevima segmenata bilo što manje diskontinuiteta u spektru signala, poželjno je da je intonacija korištenih difona međusobno što ujednačenija i neutralnija. Bolji se rezultati pri sintezi dobivaju korištenjem difona iz sredine riječi nego s krajeva, pa je važno pamtiti riječ i mjesto u riječi iz koje su izrezani. Praćenjem samo oznaka fonema u *.lab* datotekama nije moguće razabrati krajeve riječi jer se one u govoru najčešće izgovaraju zajedno, bez stanki. Postupak stoga prati i prijepis riječ po riječ, uz provjeru oznaka gdje je u snimci pogreška. Pogreške su označene oznakama <greska>, <papir>, <uzdah> itd.
Granice rezanja određene su od sredine 1. fonema do sredine 2. fonema tako da je prosječno trajanje difona jednako trajanju jednoga fonema. Kako bi se početna i krajnja vremenska oznaka difona poravnala s početkom i krajem zvučnoga zapisa, trenuci rezanja pomaknuti su od točne sredine fonema na najbližu «vanjsku» vremensku oznaku, kako je prikazano na slici 3. Oznaka T1 predstavlja interval između dviju sredina fonema, a T2 korišteni interval za izrezivanje.

using a custom Matlab function to get the occurrence frequency of each diphone in the recorded speech. Out of 1396 possible diphones, 1057 were represented with at least one instance. In the next step, the source speech segments were divided into diphones. Diphones were stored as .wav files, automatically named in the following form "1st phoneme-2nd phoneme-list_number-source_word.wav" e.g. "*a-d-11-nadajući*". In order to minimize spectral discontinuities on segment concatenation points, the prosody of the diphones used should be as constant and neutral as possible. Diphones extracted from the middles of words give better results in synthesis than those extracted from word beginnings or endings so we recorded the word and position from which the diphones were extracted. In speech, words are commonly pronounced together without pauses so it is impossible to detect word boundaries only by following phoneme labels in .lab files. For that reason the algorithm also follows the textual transcriptions word by word, also checking for labels marking errors in recordings. The errors are labeled <greska>, <papir>, <uzdah> etc. Cutting marks are set from middle of the first phoneme to the middle of the second phoneme so the average diphone duration equals the duration of a phoneme. To align the first and the diphone the pitch mark with actual sound clip beginning and end, the cutting points were moved from exact phoneme centers to the nearest "outside" pitch mark, as shown in Figure 3. Label T1 represents the range between two phoneme centers, and T2 the actual range for cutting.



Slika 3. Izrezivanje difona k-i.
Figure 3. Extraction of the diphone k-i.

Vremenske oznake dobivene su korištenjem funkcije prema [15]. Uz dobivene .wav datoteke spremaju se za svaki difon vremenske oznake, trajanje (u uzorcima), trenutak prijelaza glasa u drugi glas i pozicija u riječi iz koje je difon izrezan.
Dobiveno je ukupno više od 32 000 zapisa.
Baza je na kraju svedena na minimalni oblik, tako da je svaki difon predstavljen samo jednom inačicom. Ako je moguće, bira se difon izrezan iz sredine riječi i prosječnoga trajanja. Ako takav ne postoji, uzima se u obzir samo trajanje.
Rezultirajuća baza ima 923 unosa u .wav formatu uz pripadajuću tablicu pomoću koje se povezuju oznaka difona i ime pripadne datoteke, vremenske oznake za svaki difon i trenuci prijelaza između glasova u difonu. Veličina baze na disku je 2.47 MB.
Proces dobivanja baze u potpunosti je automatiziran, čime zamjena čitave baze novim glasom zahtijeva nakon prikupljanja snimljenoga govornog materijala i odgovarajućih fonetskih prijepisa samo ponovno pokretanje dviju funkcija *Matlab*.

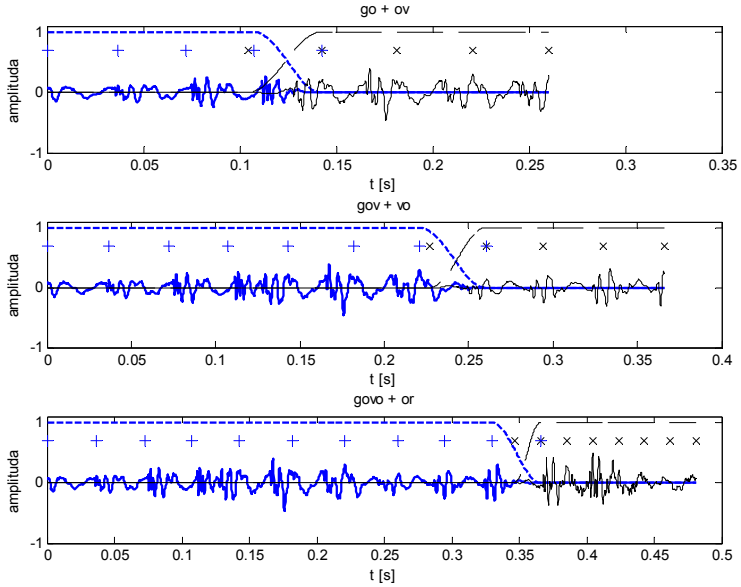
6. PRIMJENA POSTUPKA ZA TVORBU RIJEČI

Primijenjeni postupak za tvorbu riječi temelji se na algoritmu TD – PSOLA [7]. Ulazni podaci su fonetski niz koji treba pretvoriti u govor, varijable baze difona te opcionalno trajanje pauze među riječima u uzorcima. Izlazni signal dobiva se preklapanjem i zbrajanjem difona koji odgovaraju ulaznom fonetskom nizu. Dijelovi signala koji se preklapaju najprije se množe Hannovim prozorom, a dijelovi koji se ne preklapaju izravno se kopiraju u izlazni signal. Primjer spajanja difona za riječ «govor» prikazan je na slici 4:

The pitch marks were calculated using a function according to [15]. Besides the .wav file we recorded the pitch marks, duration in samples, time of transition between phonemes and the position of extraction within a word for each diphone.
A total of 32000 records were obtained.
The database was then reduced to minimal form with just one instance of each diphone. If possible the diphone was chosen from the middle of the word and of average duration. Otherwise only the duration was considered. The resulting database has 923 entries in .wav format, with corresponding table linking diphone labels and sound file, pitch marks for each diphone and times of transition between phonemes in a diphone. The database size on disk is 2.47 MB. The process of building the database is completely automated and after acquiring speech recordings and corresponding phonetic transcripts, replacing the database with a new voice requires only the running of two Matlab functions.

6. IMPLEMENTATION OF THE CONCATENATION PROCEDURE

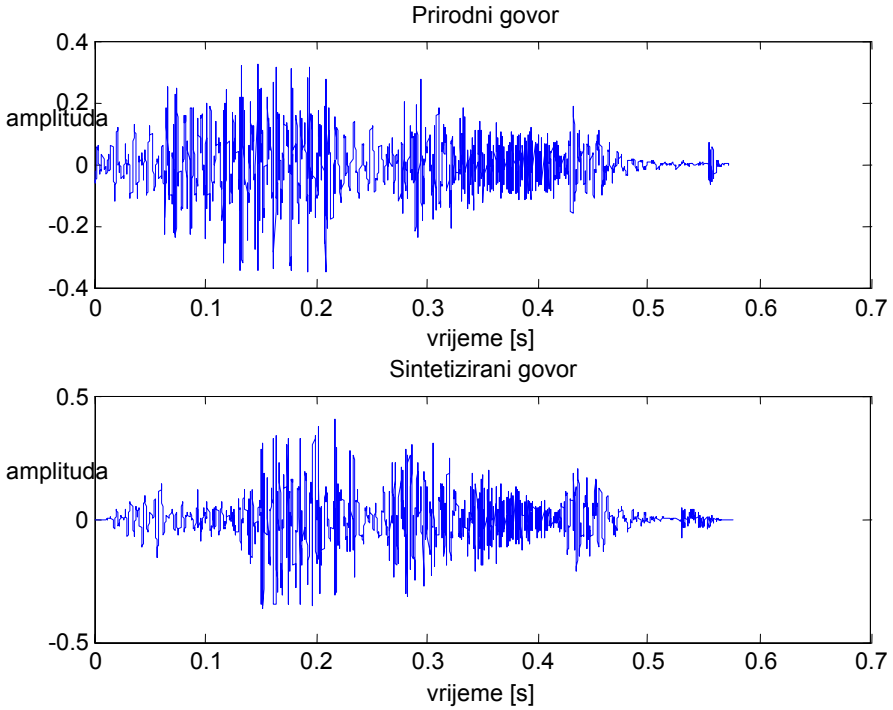
The implemented method is based on the TD-PSOLA algorithm [7]. Input data is the phonetic string to be converted to speech, the diphone database variables and optionally, the between-word pause duration in samples.
Output signal is calculated by overlapping and adding diphones accordingly to the input phonetic sequence. Overlapping parts of the signal are first multiplied by a Hanning window and the non-overlapping parts of the signal are directly copied into the output signal. Figure 4 shows the concatenation of diphones for the word "govor".



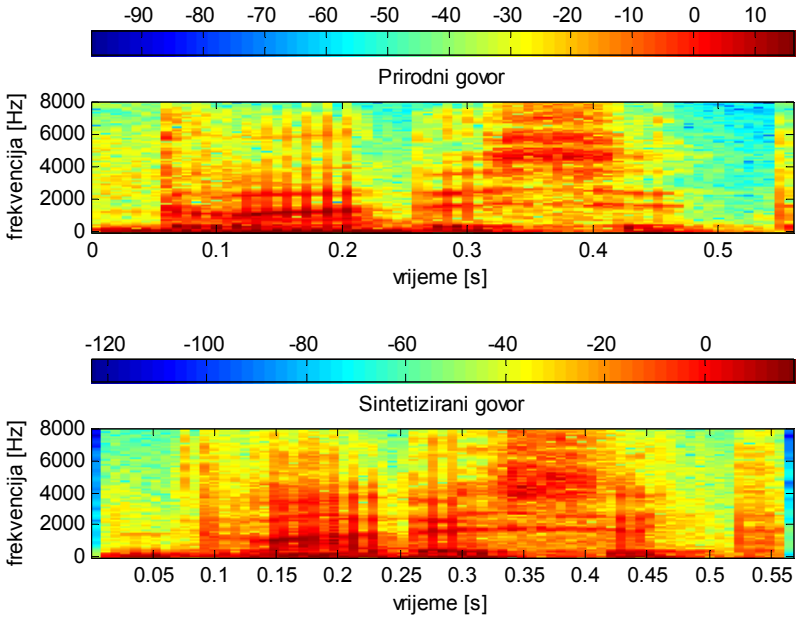
Slika 4. Tvorba riječi govor *spajanjem difona* go+ov+vo+or.
Figure 4. Synthesis of the word govor *by concatenation of diphones* go+ov+vo+or.

Isprekidanom crtom prikazane su ovojnice kojima se množi signal, punom crtom sam signal, a znakovima + i x vremenske oznake. Debljom su crtom prikazane veličine koje se odnose na izlazni signal iz prethodnoga koraka, a tanjom veličine koje se odnose na trenutačni difon. Može se uočiti poravnavanje druge vremenske oznake difona s posljednjom vremenskom oznakom postojećega signala.
Matlabovom funkcijom ufonetski.m ostvarena je osnovna funkcionalnost modula za fonetsku analizu. Ulazni argument je normalizirani tekst, a izlaz je fonetski niz. Ako tekst nije normaliziran, brojevi i simboli bit će bez promjene proslijeđeni u izlazni niz, što će sintetizator ignorirati, a kratice će biti pročitane onako kako stoje u tekstu. Pretvorba grafem-fonem vrši se jednostavnim pravilima za zamjenu znakova i korištenjem fonetskoga rječnika, prema [16].
Postupak umjetne tvorbe riječi isproban je za riječ «*dvadeset*», koja odgovara skupu sm04070105187 iz korpusa korištenoga u izradi baze. Valni oblik i spektrogram prirodnoga i umjetnog izgovora za taj slučaj prikazani su slikama 5 i 6.

The dashed line represents the multiplication envelopes, the continuous line represents the signal, and symbols + and x the pitch marks. The thicker line shows the values related to the output signal of the previous step while the thinner line shows the values related to the current diphone. Alignment of the second pitch mark of a current diphone with the last pitch mark of the existing signal is visible.
Basic functionality of the phonetic analysis module is accomplished by a Matlab function. The input argument is normalized text and the output is a phonetic string. If input text is not normalized, the numbers and symbols will be forwarded unchanged to the output stream and ignored by the synthesizer and abbreviations will be read unexpanded. Grapheme-to-phoneme translation is done by simple symbol replacement rules and using the phonetic dictionary according to [16]. The method was tested for the word "*dvadeset*" corresponding to the set sm04070105187 in the speech corpus used to build the database. Waveform and spectrogram of natural and synthesized speech of that word is shown in figures 5 and 6.



Slika 5. Usporedba valnih oblika prirodnoga i sintetiziranoga izgovora sm04070105187.
Figure 5. Waveform comparison of natural and synthesized speech, utterance sm04070105187.



Slika 6. Usporedba spektrograma prirodnoga i sintetiziranoga izgovora sm04070105187.
Figure 6. Spectrogram comparison of natural and synthesized speech, utterance sm04070105187.

7. OCJENA KVALITETE SUSTAVA ZA UMJETNU TVORBU GOVORA

Za utvrđivanje kvalitete ostvarenoga sustava za umjetnu tvorbu govora provedena je anketa među 19 ispitanika. Anketirane osobe najprije su preslušale ulomak sintetiziranoga govora, a zatim su odgovarale na dvije skupine pitanja. Prvi skup pitanja odnosio se na ocjenu kvalitete sinteze, dok se drugi skup odnosio na provjeru razumljivosti. Na ljestvici od pet stupnjeva ocjenjivani su kvaliteta sintetiziranoga govora, razumljivost, prirodnost te učestalost pojave nepravilnosti pri izgovoru. Ispitanici su tada odgovorili na pitanje smatraju li da je sintetizirani govor primjeren za davanje odgovora u telefonskom sustavu za automatsko posredovanje informacija, gdje su ponuđena tri odgovora: da / da, uz popravke / ne. Odgovori na drugi skup pitanja bili su sadržani u izgovorenoj poruci, čime se provjeravala razumljivost. Sadržaj ulomka bila je vremenska prognoza u trajanju od 41s, a trajanje stanki između riječi je 1200 uzoraka. Bilježili su se ovi podaci o ispitanicima: dob, spol, stupanj obrazovanja, struka (informatika, lingvistika ili drugo), jesu li imali prijašnjih kontakata sa sintetizatorima govora i kakvim te jesu li prije već sudjelovali u evaluaciji.

7.1. Rezultati ispitivanja

Za prvi skup pitanja određene su prosječne ocjene svih ispitanika. Dobiveni rezultati prikazani su na slici 7. Odgovor na pitanje o primjerenosti upotrebe

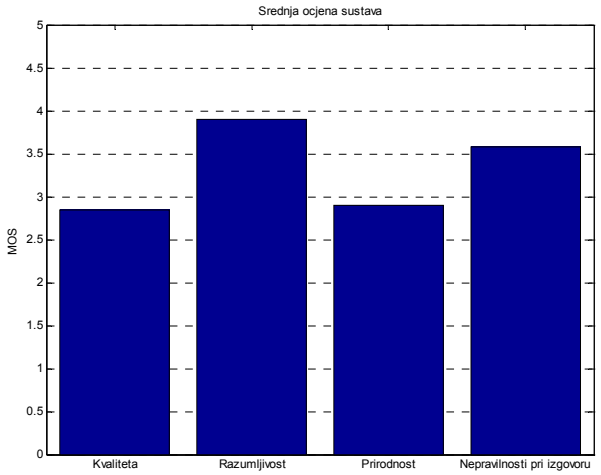
7. TTS SYSTEM QUALITY ASSESMENT

To assess the quality of the implemented speech synthesis system, we conducted a survey among 19 subjects. The subjects first listened to a fragment of synthesized speech and then answered two sets of questions. The first set referred to the quality of synthesized speech while the second was used to test the intelligibility. Rated on a scale from one to five were quality of synthesized speech, intelligibility, naturalness and frequency of errors in pronunciation. Subjects then answered if they believed the synthesized speech was appropriate for usage in automated telephone information service, with three options: yes, yes with improvements, no. The answers to the second set of questions were contained in the spoken message and were used to test intelligibility. Content of the fragment was a weather report, 41 seconds in duration and the pause between words was 1200 samples. We also recorded the subjects' age, sex, occupation, profession (information science, linguistics, other), whether they had previous contact with speech synthesis and how and if they had previously participated in the survey.

7.1 Survey Results

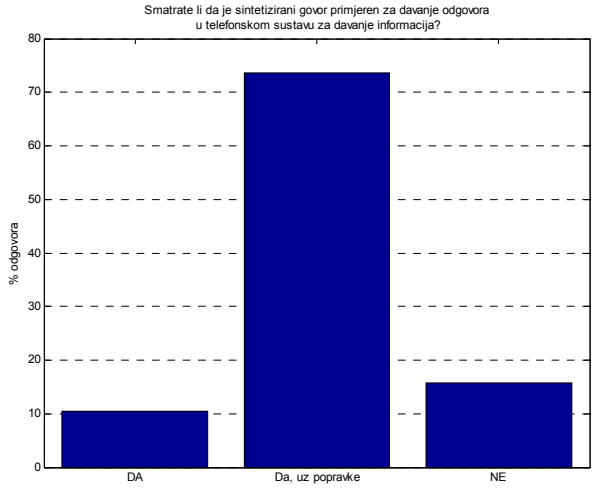
For the first set of questions we calculated the average ratings from all subjects. The results are shown in Figure 7. The answers to the question about the appropriateness

sintetiziranoga govora u telefonskim sustavima za davanje informacija prikazan je grafički na slici 8.



Slika 7. Prosječna ocjena (MOS) sustava.
Figure 7. System mean opinion score (MOS)

of synthesized speech in telephone information systems are shown graphically in Figure 8.



Slika 8. Ocjena prikladnosti sintetiziranoga govora u sustavu pružanja informacija.
Figure 8. The mean rate of appropriateness for TTS usage in a spoken dialog system.

Na pitanja koja su se odnosila na razumljivost bilo je 76% točnih odgovora.
Najviše ispitanika (petero) kao primjedbu je napisalo da izgovor nije dovoljno tečan kao u prirodnom govoru, tj. da govornik previše zastajkuje. Više zamjerki bilo je i na razumljivost dijela prognoze (fraza «*Na Jadranu pretežito u unutrašnjosti djelomice sunčano.*»).
Razumljivost dobivenoga govora bitno ovisi o kvaliteti korištenih difona u bazi. Pri izradi baze korišten je automatski postupak određivanja vremenskoga segmenta svakoga difona. Za poboljšanje razumljivosti trebalo bi ručno odrediti intervale trajanja difona.
Veća kvaliteta i prirodnost mogle bi se postići dinamičkim izborom difona pri njihovu spajanju. U tom slučaju baza ne bi imala samo jednu instancu svakoga difona, već bi se zadržale sve ili veći broj dostupnih instanci. Pri ulančavanju segmenata od dostupnih inačica odgovarajućega difona odabrat će se ona koja će dati najmanja prozodijska i spektralna izobličenja u odnosu na prethodni segment [10]. Alternativni je pristup korištenje modifikacije valnoga oblika segmenata metodom TD-PSOLA kako je opisano u [7], čime se osnovna frekvencija glasa na mjestima spoja difona može izjednačiti i time smanjiti nastala izobličenja. Taj postupak nudi i druge prednosti, kao što je mogućnost promjene brzine izgovora i visine glasa te prilagodba željenoj prozodiji. Modifikacija valnog oblika međutim može i narušiti prirodnost izgovora.

There were 76% correct answers to questions that tested intelligibility.
The most complaints (five) were that the speech was not as fluent as natural, i.e. the speaker is stuttering. In addition, there were some complaints about the intelligibility of a part of the weather report (phrase «*Na Jadranu pretežito u unutrašnjosti djelomice sunčano.*»).
Intelligibility of the synthesized speech greatly depends on the quality of diphones in the database. In database construction, we used an automated process of diphone segementation without verification. To improve intelligibility, diphones of lesser quality should be manually segmented.
Better quality and naturalness could be achieved with dynamic selection of diphones at the time of concatenation. In that case, the database would not have just one instance of each diphone, instead a number of or all available instances would be kept. In segment concatenation from the available instances of a matching diphone, the one with the least resulting prosodic and spectral distortion with regard to the previous segment would be chosen [10].
The alternative approach is to use waveform modification using the PSOLA technique, as described in [7], equalizing the fundamental pitch between segments in concatenation points and reducing distortion. This approach has other advantages, such as changing the rate (speed) of speech and adapting to the desired prosody. Waveform modification can however impair the perceived naturalness of speech.

8. ZAKLJUČAK

U ovome radu opisan je sustav za umjetnu tvorbu hrvatskoga govora. Prikazane su struktura općenitog sustava kao i moguće metode za sintezu govora. Pobliže je opisana metoda ulančavanjem manjih akustičnih segmenata – difona, na temelju koje je realiziran sustav u programskom alatu *Matlab* te izrada baze difona za potrebe toga sustava.
Za provjeru kvalitete orimijenjenog postupka i baze provedeno je ispitivanje među 19 osoba koje su ocijenile kvalitetu, razumljivost, prirodnost i učestalost pogrešaka kod umjetno tvorenoga govora relativno visokom srednjom ocjenom. Prikazani su rezultati ispitivanja i predloženi mogući postupci za poboljšanje kvalitete dobivenoga govora.

8. CONCLUSION

This paper describes a system for Croatian speech synthesis. The structure of a general text-to-speech system is shown, as well as possible synthesis methods. In more detail, the diphone concatenation method is shown, which was used in the system built using the Matlab tool, as well as the diphone database. For quality assessment of the resulting system and database, a survey was conducted among 19 subjects rating quality, intelligibility, naturalness and frequency of errors in pronunciation in synthesized speech. Relatively high mean opinion rates from the survey are presented and some possible procedures for improving the quality of speech were suggested.

LITERATURA
REFERENCES

[1] Taylor, P., Black, A., Caley, R.: *The architecture of the festival speech synthesis system*, The Third ESCA Workshop in Speech Synthesis, Jenolan Caves, Australia, 1998, pp 147-151.

[2] Clark, R. A.J., Richmond, K., King S.: *Festival 2 - build your own general purpose unit selection speech synthesiser*, Proc. 5th ISCA workshop on speech synthesis, 2004.

[3] Dutoit, T.: *An Introduction to Text-To-Speech Synthesis,* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.

[4] Bakran, J., Lazić, N.: *Fonetski problemi difonske sinteze hrvatskoga govora*, Govor XV, 1998. 2, pp. 103-116.

[5] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T.: *Speech parameter generation algorithms for HMM-based speech synthesis*, Proc. of ICASSP, June 2000, pp.1315-1318

[6] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A.W., Tokuda, K.: *The HMM-based speech synthesis system version 2.0*, Proc. of ISCA SSW6, Bonn, Germany, Aug. 2007.

[7] Moulines, E., Charpentier, F.: *Pitch Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones*, Speech Communication, 1990, 9:453-467.

[8] Sečujski, M., Obradović, R., Pekar, D., Jovanov, Lj., Delić, V.: *AlfaNum sistem za sintezu govora na osnovu teksta na srpskom jeziku (engleski)*, TSD 2002, Brno, 2002, pp. 237-244

[9] Gros, J., Pavešić, N., Mihelič, F.: *Text-to-Speech Synthesis: A Complete System for the Slovenian Language,* Journal of Computing and Information Technology – CIT 5, 1997, 1, 11-19.

[10] Huang, X., Acero, A., Hon, H.: *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, Prentice Hall, Upper Saddle River, New Jersey, 2000.

[11] Klatt D. H., *"Review of Text to Speech Conversion for English,"* Journal of the Acoustic Society of America, vol. 82, pp. 737-793, 1987.

[12] Martinčić-Ipšić, S., Ipšić, I.: *Croatian HMM-based Speech Synthesis*, Journal of Computing and Information Technology – CIT 14, 2006, Vol. 4, pp. 307-313.

[13] Rabiner, L.R., Schafer, R.W.: *Digital Processing of Speech Signals*, Prentice Hall, 1978.

[14] Martinčić-Ipšić, S., Matešić, M., Ipšić, I.: *Korpus hrvatskoga govora*, Govor : časopis za fonetiku. I (2004), 2; 135-150.

[15] Goncharoff, V., Gries, P.: *An algorithm for accurately marking pitch pulses in speech signals*, IASTED International conference SIP '98, Nevada, USA

[16] Bakran, J., Horga, D.: *SAMPA for Croatian.* Govor. XIII. Vol.1-2. (1996) p. 99-104.

Primljeno / Received: 30.4.2008

Prihvaćeno / Accepted: 3.9.2008

Izvornoznanstveni članak

Original scientific paper

Adresa autora / Authors' address:
Miran Pobar
Sanda Martinčić-Ipšić
Sveučilište u Rijeci, Odjel za informatiku

Omladinska 14
51000 Rijeka
HRVATSKA
mpobar@ffri.hr
smarti@ffri.hr
Ivo Ipšić
Sveučilište u Rijeci, Tehnički fakultet
Vukovarska 58
51000 Rijeka
HRVATSKA
ivoi@ffri.hr