The Framework Of The Turkish Syllable-Based Concatenative Text-To-Speech System With Exceptional Case Handling

ZEYNEP ORHAN, ZELIHA GÖRMEZ
Computer Engineering Department
Fatih University
34500, Istanbul
TURKEY
zorhan@fatih.edu.tr, zcetin@fatih.edu.tr

Abstract: - This paper describes the TTTS (Turkish Text-To-Speech) synthesis system, developed at Fatih University for Turkish language. The framework of the Turkish syllable-based concatenative text-to-speech system with exceptional case handling is introduced. TTTS is a concatenative TTS system aiming to advance the process of developing natural and human sounding Turkish voices. The resulting system is implemented by concatenating pieces of pre-recorded limited number of speech units that are stored in a database. Systems differ in the size of the stored speech units affecting the output range, the quality and the clarity, therefore the number of the concatenation units, synthetic units obtained and the computational power required should be kept in balance. The letters of the Turkish alphabet and the syllables that consist of two letters at most are used as the smallest phonemes in the context of this study. The syllables that have more than two letters are derived from these smallest units. The words, which are generally borrowed from other languages throughout cultural interactions, present exceptional behaviors and should be handled specifically. The results are evaluated by using the Degradation Mean Opinion Score (DMOS) method.

Key-Words: - Text-to-speech (TTS), Speech Synthesis, Concatenative Turkish TTS .

1. Introduction

The aim of the speech synthesis is producing the human speech artificially either in software or hardware. The natural language text is converted into speech by text-to-speech (TTS) systems.

These systems have been widely used as assistive technological tools for a long time. The pre-school kids, the people who have visual impairments or reading disabilities, and the ones who suffer from severe speech impairment can benefit from these systems. The news web sites that convert written news to audio content, entertainment productions such as games, cartoons, mobile tools, preparation of audio supplementary materials in various fields, automated question-answering systems, attaining certain information (price list, the weather forecasting report, etc.) and vocalizing e-mail, fax, sms, and daily journals for handicapped ones are only a limited number of items that can be listed as the typical application areas of TTS. The aim of the TTS is that the system converts all digital texts and printed texts via OCR (Optical character recognition) into speech automatically. Commercial and non-commercial systems have been continuously developed and recent advances are promising for future applications. The processes that take place in a TTS system is shown as a block diagram in Fig. 1.

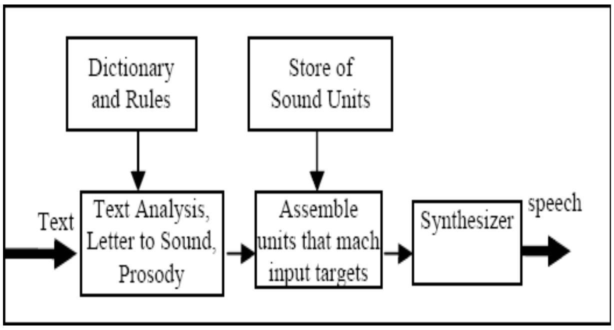

Dictionary and Rules
Store of Sound Units
Text
Text Analysis, Letter to Sound, Prosody
Assemble units that mach input targets
Synthesizer
speech

Fig. 1: Block Diagram of the TTS System [6]

TTS systems can be categorized into two groups as the ones that are using formant synthesis and concatenative synthesis[1]. The strengths and weaknesses of both technologies exist depending on the requirements of the systems where they are employed.

The concatenative approach is based on the small pieces of recorded speech. In this approach, to prepare "speech database", the small pieces are

[1] There are some other approaches such as articulatory synthesis but not considered in the context of this study.

either cut from the recordings or recorded directly and then stored. Then, at the synthesis phase, units selected from the speech database are concatenated and, the resulting speech signal is synthesized as output. In this approach, the longer the phoneme means the more success of the system. Concatenative synthesis has the potential for producing the most natural-sounding synthesized speech. However, differences between natural variations in speech and the automatic segmentation of the waveforms can cause audible glitches in the output [8] .

Unit selection is one of the concatenative approaches. It uses large speech database and applies only a small amount of digital signal processing (DSP) to the recorded speech providing the greatest naturalness. On the other hand, the size of the database required and selecting the appropriate unit from this large database can cause problems in this approach [10] .

Another concatenative approach is called the diphone synthesis that uses a minimal speech database containing all possible diphones (sound-to-sound transitions) in a language. The size of the diphone database may vary depending on the language. These units are combined by DSP techniques in the synthesis process resulting in a quality less than the unit synthesis but generally better than the formant synthesis and keeping the size of the database small [8] .

The alternative to the concatenative synthesis is formant synthesis which synthesizes artificial, robotic sound speech by using an acoustic model instead of human speech and avoids the acoustics glitches. The memory and microprocessor power requirement is less than the aforementioned techniques; therefore it is suitable for the limited devices [8] .

In this study, the concatenative system is preferred since Turkish is an agglutinative language that is very productive and it is likely to derive plenty of new words by adding affixes to words by adding suffixes. The idea of keeping a database of all possible combinations will be burdensome and inconvenient. It will yield a very large database ranging into gigabytes and will require frequent modifications. Therefore, concatenative approach, more specifically an approach similar to the diphone synthesis is chosen for Turkish to keep the size of the database small and to have reasonable amount of DSP. Similar approaches are used for other languages in TTS applications [2]

In the next section, the speech database requirements are explained. The third section provides the details of the implementation and the fourth section summarizes testing environment and its results. The improvements than can be studied in the future and some concluding remarks are presented in the fifth chapter.

2. Speech Database Preparation

There are 21 consonants (C) and 8 vowels (V), yielding a total of 29 characters in Turkish alphabet. The syllables in Turkish are formed with the combination of consonants and vowels in many ways. Syllables are generally formed from 1-6 characters and contain a vowel and consonants with some minor exceptions. However, some of these, especially the syllables that has 5 or 6 characters are very rare.

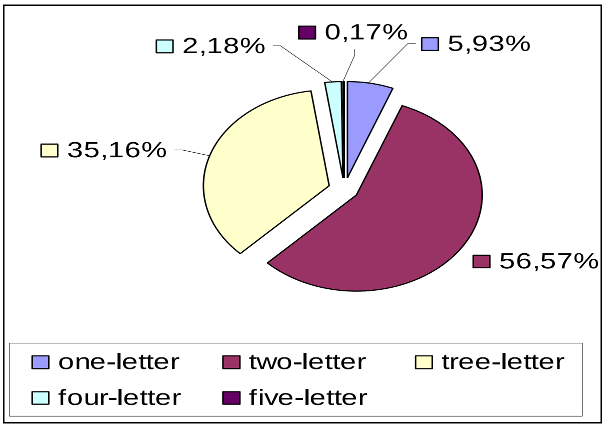


Fig. 2: The percentages of various syllable lengths in Turkish

The percentages of these syllables are given in Fig. 2 that is obtained from the Turkish corpora prepared in a research of this domain [1] . The ratios of the syllables support the claim asserted above. Therefore, only the six types of syllables in Turkish as shown in Table 1 are considered. The speech database of Turkish includes single or double letter sounds as the smallest phoneme in our system. The rest of the syllables are formed from the concatenation of these sounds. The longer syllables were synthesized as follows:

CV+C → CVC
VC+C → VCC
CC+V → CCV
CV+C+C → CVCC

Single letter syllables can have only the vowels and double letter syllables may have one consonant and one vowel and their order may change. The syllables that have 3 or more characters can be obtained by adding a consonant to double letter syllable. The resulting database has 365 recordings as a total and Table 2 shows how this is obtained. The synthesis system tested by making use of the method Degradation Mean Opinion Score (DMOS) [5]

Firstly, it was planned to process the speech records of Turkish instead of forming new voice file records for the syllables, cutting and saving the syllables as new files. But, we deemed it suitable to save the syllables, since it is hard to determine the syllable boundaries in words. The software called Speech Analyzer of SIL International Organization is used for speech processing [7] . The possible CV-VC combinations for Turkish are shown in the Table 3. Sound files are kept with the same name given in the table.

Table 1: The structure of Turkish syllables and required speech records

Syllable structure	Sample syllables	Required	Total
V	a, e, ı, i, o, ö, u, ü	29	29
VC	ab, ac, aç, ad, … ,az, eb, ec,…	8*21	168
CV	ba, be, bı, bi,…, za, ze, zı, zi, …	8*21	168
CVC	bak, git, say, kır, …	21*8*21	3528
VCC	ast, üst, ırk, …	8*21*21	3528
CCV	tra, pla, tre	21*21*8	3528
CVCC	Türk, kürt, sırt,	21*8*21*21	74088
Total			**85037**

Table 2:Diphones kept in Turkish speech database

Syllable Combinations	Possible values
CV	21*8
VC	21*8
V	8
C	21
Total	**365**

Table 3 : CV-VC Syllable Combinations

C\V	a		e		ı		i		u		ü		o		ö	
b	ab	ba	eb	be	ıb	bı	ib	bi	ub	bu	üb	bü	ob	bo	öb	bö
c	ac	ca	ec	ce	ıc	cı	ic	ci	uc	cu	üc	cü	oc	co	öc	cö
ç	aç	ça	eç	çe	ıç	çı	iç	çi	uç	çu	üç	çü	oç	ço	öç	çö
d	ad	da	ed	de	ıd	dı	id	di	ud	du	üd	dü	od	do	öd	dö
f	af	fa	ef	fe	ıf	fı	if	fi	uf	fu	üf	fü	of	fo	öf	fö
g	ag	ga	eg	ge	ıg	gı	ig	gi	ug	gu	üg	gü	og	go	ög	gö
ğ	ağ	ğa	eğ	ğe	ığ	ğı	iğ	ği	uğ	ğu	üğ	ğü	oğ	ğo	öğ	ğö
h	ah	ha	eh	he	ıh	hı	ih	hi	uh	hu	üh	hü	oh	ho	öh	hö
j	aj	ja	ej	je	ıj	jı	ij	ji	uj	ju	üj	jü	oj	jo	öj	jö
k	ak	ka	ek	ke	ık	kı	ik	ki	uk	ku	ük	kü	ok	ko	ök	kö
l	al	la	el	le	ıl	lı	il	li	ul	lu	ül	lü	ol	lo	öl	lö
m	am	ma	em	me	ım	mı	im	mi	um	mu	üm	mü	om	mo	öm	mö
n	an	na	en	ne	ın	nı	in	ni	un	nu	ün	nü	on	no	ön	nö
p	ap	pa	ep	pe	ıp	pı	ip	pi	up	pu	üp	pü	op	po	öp	pö
r	ar	ra	er	re	ır	rı	ir	ri	ur	ru	ür	rü	or	ro	ör	rö
s	as	sa	es	se	ıs	sı	is	si	us	su	üs	sü	os	so	ös	sö
ş	aş	şa	eş	şe	ış	şı	iş	şi	uş	şu	üş	şü	oş	şo	öş	şö
t	at	ta	et	te	ıt	tı	it	ti	ut	tu	üt	tü	ot	to	öt	tö
v	av	va	ev	ve	ıv	vı	iv	vi	uv	vu	üv	vü	ov	vo	öv	vö
y	ay	ya	ey	ye	ıy	yı	iy	yi	uy	yu	üy	yü	oy	yo	öy	yö
z	az	za	ez	ze	ız	zı	iz	zi	uz	zu	üz	zü	oz	zo	öz	zö

3. Text-to-Speech Method

The first step of TTS system is preprocessing. The preprocessing of the text to be converted into speech requires the followings: Clearing all unnecessary format information, normalizing the text by using all upper/lowercase letters, and converting all white spaces (\t, \n, \s) into single space characters. This step is followed by the syllabification phase. The Zemberek[2] Project [4] is used as a tool for this step in our system. The syllables that have more than two letters are derived from the smallest units. The words, which are generally borrowed from other languages throughout cultural interactions, present exceptional behaviors and should be handled specifically.

Table 4: Special Characters and their vocalizations

Character	Vocalization (alternatives are seperated by /)
@	Et
%	Yüzde
&	Ve
|	Veya
#	Diyez
*	Çarpı
/	Bölü
-	Eksi/Tire
+	Artı
>	Büyüktür
<	Küçüktür
(	Aç parantez
)	Kapa parantez
=	Eşittir
~	Yaklaşık
€	Yuro
$	Dolar
_	Alt çizgi
,	Virgül (for decimal number)/None
.	None
:	None
;	None
?	None
!	None
'	None
"	None

3.1. Preprocessing and Handling Exceptions

The following steps are used for text preprocessing

- The liaison is checked before removing formatting instructions. The liaison is the grammatical circumstance in which a usually silent consonant at the end of a word is pronounced at the beginning of the word that follows it. The pattern that has a potential for being a liaison is searched. This pattern generally has the form of "C V", i.e. "ConsonantSpaceVowel". The matched patterns are replaced by "CV", in other words the space is deleted and the last part of the first word is combined with the first part of the second word. In this way they are treated as a single syllable. The examples of the liaison are given by → in the following lines of a Turkish poem:

  "Dönülmez→akşamın ufkundayız vakit çok geç
  Bu son fasıldır→ey ömrüm nasıl geçersen geç"

  However, the following example is not a liaison, since the comma disturbs the rule. Therefore, the punctuations are required for liaison detection.

  Annem,→ablam geldi.(Liaison is not applicable at → point).

- The special characters are controlled by using a list given in Table 4. The special characters that require vocalization are replaced by "SpaceSpecialCharacterSpace" pattern and treated as a single word whose synthesis will be done according to the second column of the table. The ones that have *None* in the corresponding column are replaced by space only. Some of them are ambiguous and have more than one possible vocalization depending on the context. Comma(,) is vocalized in the context of a number, but not in the context of a normal word.
- After the previous pattern matching and replacement operations extra whitespaces are no more required and can be removed safely. Only single spaces are left for the word boundary detection.
- The text is divided into word pieces by considering the word boundaries and kept as an array. Then these are sent to the syllabification phase and other exceptional cases are handled there.

[2] Zemberek is an open source, platform independent, general purpose Natural Language Processing library and toolkit designed for Turkic languages, especially Turkish. [4]

Table 5: Possible combinations of two consecutive consonants at the end and beginning of a syllable in Turkish (1=Possible, 0=Impossible or ignored)

Position=End of syllable																					
	b	**c**	**ç**	**d**	**f**	**g**	**ğ**	**h**	**j**	**k**	**l**	**m**	**n**	**p**	**r**	**s**	**ş**	**t**	**v**	**y**	**z**
b	0	0	0	**1**	0	0	0	0	0	0	**1**	0	0	0	0	0	0	**1**	0	0	0
c	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ç	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0
f	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	**1**	0	0	0
g	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ğ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
h	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0
j	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	**1**	0	0	0
l	0	0	**1**	0	**1**	**1**	0	**1**	0	**1**	0	**1**	0	**1**	0	**1**	0	**1**	0	0	0
m	**1**	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	**1**	0	0	0
n	0	0	**1**	0	**1**	**1**	0	0	0	**1**	0	0	0	0	0	**1**	0	**1**	0	0	**1**
p	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	0	**1**	**1**	**1**	**1**	0	**1**	**1**	**1**	0	**1**	**1**	**1**	0	**1**	**1**	**1**	**1**	0	**1**
s	0	0	0	0	0	0	0	0	0	**1**	0	0	0	**1**	**1**	0	0	**1**	0	0	0
ş	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	**1**	0	0	0
t	0	0	0	0	**1**	0	0	0	0	0	0	**1**	0	0	**1**	0	0	0	0	0	0
v	0	0	0	0	0	0	0	0	0	**1**	0	0	0	**1**	0	**1**	0	**1**	0	0	0
y	0	0	0	**1**	0	0	0	**1**	0	0	**1**	0	**1**	**1**	**1**	**1**	**1**	**1**	0	0	0
z	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0

Position=Beginning of syllable																					
	b	**c**	**ç**	**d**	**f**	**g**	**ğ**	**h**	**j**	**k**	**l**	**m**	**n**	**p**	**r**	**s**	**ş**	**t**	**v**	**y**	**z**
b	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0
c	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ç	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0
f	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	**1**	0	0	0	0	0	0
g	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	**1**	0	0	0	0	0	0
ğ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
h	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0
j	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	**1**	0	0	0	0	0	0
l	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
m	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	**1**	**1**	0	0	0	0	0
r	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s	0	0	0	0	**1**	0	0	0	0	**1**	**1**	**1**	0	**1**	0	0	0	**1**	0	0	0
ş	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
t	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0
v	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 6:Examples of possible combinations of two consecutive consonants at the end and beginning of a syllable in Turkish(N/A=Not Applicable)

Consec.	Example	English	Example	English
Cons.	**(Pos.=End)**	**Translation**	**(Pos.=Begin)**	**Translation**
bd	abd	slave,servant	N/A	N/A
bl	fabl	fable	N/A	N/A
br	N/A	N/A	branş	branch
bt	zabt	restrain	N/A	N/A
cd	vecd	entrancement	N/A	N/A
dh	medh	praise	N/A	N/A
dr	N/A	N/A	dram	drama
fl	N/A	N/A	flüt	flute
fr	N/A	N/A	fren	brake
fs	nefs	essence,flesh	N/A	N/A
ft	çift	double,pair	N/A	N/A
gl	N/A	N/A	gladyatör	gladiator
gr	N/A	N/A	gram	gram
hd	ahd	vow	N/A	N/A
hr	N/A	N/A	hristiyan	christian
ht	baht	luck	N/A	N/A
kl	N/A	N/A	kla-sik	classic
kr	N/A	N/A	krem	cream
ks	lüks	luxury	N/A	N/A
kt	dir-rekt	direct	N/A	N/A
lç	felç	apoplexy	N/A	N/A
lf	golf	golf	N/A	N/A
lg	telg-raf	telegram	N/A	N/A
lh	sulh	peace	N/A	N/A
lk	ilk	first	N/A	N/A
lm	film	film	N/A	N/A
lp	kulp	handle	N/A	N/A
ls	vals	waltz	N/A	N/A
lt	alt	bottom	N/A	N/A
mb	amb-lem	emblem	N/A	N/A
mp	komp-leks	complex	N/A	N/A
mt	semt	district	N/A	N/A
nç	genç	young	N/A	N/A
nf	enf-las-yon	inflation	N/A	N/A
ng	mi-ting	meeting	N/A	N/A
nk	renk	color	N/A	N/A
ns	fi-nans	finance	N/A	N/A
nt	hint-li	indian	N/A	N/A
nz	bronz	bronze	N/A	N/A
pl	N/A	N/A	plan	plan
pr	N/A	N/A	pro-fe-sör	professor
ps	N/A	N/A	psi-ko-log	psychologist
rç	borç	debt	N/A	N/A
rd	ard	consecutive	N/A	N/A
rf	harf	letter	N/A	N/A
rg	morg	morgue	N/A	N/A
rh	zırh	armor	N/A	N/A
rj	şarj	charge	N/A	N/A
rk	ırk	race	N/A	N/A
rm	form	form	N/A	N/A

Consec. Cons.	Example (Pos.=End)	English Translation	Example (Pos.=Begin)	English Translation
rn	mo-dern	modern	N/A	N/A
rp	sarp	steep	N/A	N/A
rs	ders	lesson	N/A	N/A
rş	marş	march,anthem	N/A	N/A
rt	kort	court	N/A	N/A
rv	re-zerv	reserve	N/A	N/A
rz	arz	earth,supply	N/A	N/A
sf	N/A	N/A	sfenks	sphinx
sk	kask	helmet	skandal	scandal
sl	N/A	N/A	slayt	slide
sm	N/A	N/A	smo-kin	tuxedo
sp	esp-ri	witticism	spor	sports
sr	N/A	N/A	N/A	N/A
st	ast	junior	stres	stress
şk	aşk	amour,love	N/A	N/A
şt	ser-gü-zeşt	adventure	N/A	N/A
tf	lutf	grace	N/A	N/A
tm	ritm	rhythm	N/A	N/A
tr	fötr	felt	tren	train
vk	zevk	enjoyment	N/A	N/A
vr	sevr	Sèvres,ox	N/A	N/A
vs	Dos-to-yevs-k	Dostoyevsky	N/A	N/A
vt	lo-kavt	lockout	N/A	N/A
yd	N/A	N/A	N/A	N/A
yh	a-leyh	against	N/A	N/A
yl	kok-teyl	cocktail	N/A	N/A
yn	di-zayn	design	N/A	N/A
yp	teyp	tape player	N/A	N/A
yr	seyr	wtach	N/A	N/A
ys	ays-berg	iceberg	N/A	N/A
yş	N/A	N/A	N/A	N/A
yt	la-kayt	uninterested	N/A	N/A
zm	fe-mi-nizm	feminism	N/A	N/A

3.2. Obtaining Concatenation Units

Firstly, the word alignment of the given text is achieved by presuming the spaces as word boundaries. Later on, each word was syllabified by Zemberek and output of it is considered for further processing for the syllables that are longer than two characters. Single or double letter syllables are directly converted to speech by using the prerecorded speech files; however the longer ones are spitted into single or double letter syllables and are synthesized as follows:

CV+C (dı-r)
CV+C+C (tü-r-k)

Two consecutive consonants may be an indicator of an exceptional case. Possible combinations and their examples are provided in Table 5 and Table 6. These tables are used for the decision of the consecutive consonants synthesis. The following steps are taken during this phase depending on the length of the syllables and exceptional cases:

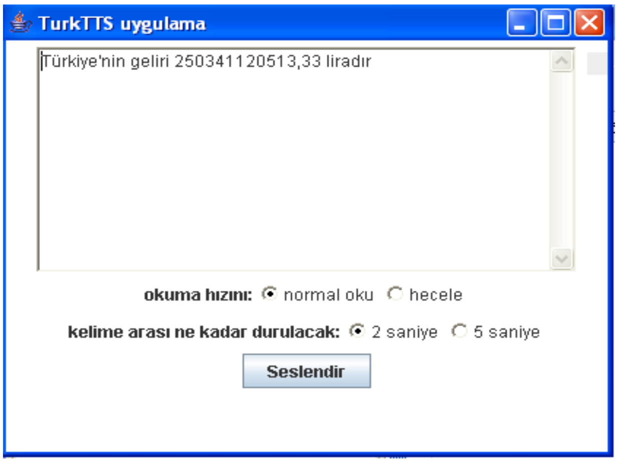
- If the length is 1 then
  - If it is a special character, then the related sound from Table 4 is chosen.
  - If it is a vowel, then it is pronounced as it is.
  - If it is a consonant, then it is pronounced by concatenating an *e* or *a* sound to it.
- If the length is 2 or more then
  - The epenthesis is checked. If required additional sounds are inserted by using some heuristics and exceptional rules. Examples are the word *grip(flu)* changed to *gırip*, *profesör(professor)* to *purofesör* etc.
  - It is sent to Zemberek for syllabification

- If it fails then the word is accepted as an abbreviation and pronounced as word by word.
- If it succeeds and the syllable includes two consecutive consonants, Table 5 is checked for the valid combination. If it is valid, it is synthesized normally; otherwise it is again accepted an abbreviation. The word *fabl (fable)* has no syllables and includes two consonants at the end and synthesized as fa+b+l. On the other hand, *ABS (Anti-lock Braking System)* has no syllables and ending with two consonants which is not a valid combination and accepted as an abbreviation.

There are many types of ambiguities occurring in the TTS systems. One such kind of an ambiguity is related to the numbers. For example, if the number in textual form is 8540178, then it is converted into speech by different ways depending on the context. If it is a phone number, it is pronounced as 854 01 78 (eight hundred fifty four zero one seventy eight), on the other hand, if it is a currency, than is it is converted as 8 540 178 (eight million five hundred forty thousand one hundred seventy eight). In the TTTS, all numbers are converted in the latter form for the time being. Besides, it can vocalize the decimal numbers. The system provides some various functionalities for TTS. The level of the reading speed can be justified depending on the usage of the system and the level of the reader by using the provided options in the user interface. The system interface is given in Fig. 3.

4 Testing the System

There are some basic criteria for measuring the performance of a TTS system: These are the similarity to the human voice (naturalness) and the ability to be understood (intelligibility). The ideal speech synthesizer is both natural and intelligible, or at least try to maximize both characteristics. Therefore, the aim of TTTS is also determined as to synthesize the speeches in accordance with natural human speech and clarify the sounds as much as possible. The study is tested by making use of the DMOS. The MOS [1] is expressed as a single number in the range 1 to 5, where 1 is lowest perceived quality, and 5 is the highest perceived quality. MOS tests for voice are specified by ITU-T recommendation P.800. The MOS is generated by averaging the results of a set of standard, subjective tests where a number of listeners rate the perceived audio quality of test sentences read aloud by both male and female speakers over the communications medium being tested. A listener is required to give each sentence a rating using the rating scheme in Table 7. The perceptual score of the method DMOS is calculated by taking the mean of the all scores of each sentence.



Türkiye'nin geliri 250341120513,33 liradır

okuma hızını: ● normal oku ○ hecele

kelime arası ne kadar durulacak: ● 2 saniye ○ 5 saniye

Fig. 3:TurkTTS User Interface

Table 7: Mean Opinion Score (MOS)

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

The users are given a few sentences and asked to grade by using a number in the range of 1-5 for each word, sentence, number and special character to test the system. The results are shown in Table 8 and depicted in Fig. 4. The mean of these values is computed as the DMOS value which is equal to 3.7. The results assert the following conclusions: The words including two-letter syllables got higher scores since the syllables of these words are already recorded. For the words including three-letter syllables, the concatenation of the syllables of two audios led to a decline in the score. The decimal number with comma enabled it to get a high score. Special characters and numbers have been rated better as expected, since the quality of the sound depends on the unit length used in the database.

Table 8: DMOS scores of the system

	Excellent(5)	Good(4)	Fair(3)	Poor(2)	Bad(1)	Avg.
Words		4		1		3,6
Sentence		1	3	1		3
Number	3	2				4,6
Special char.	1	1	3			3,6
System's overall performance average						3,7

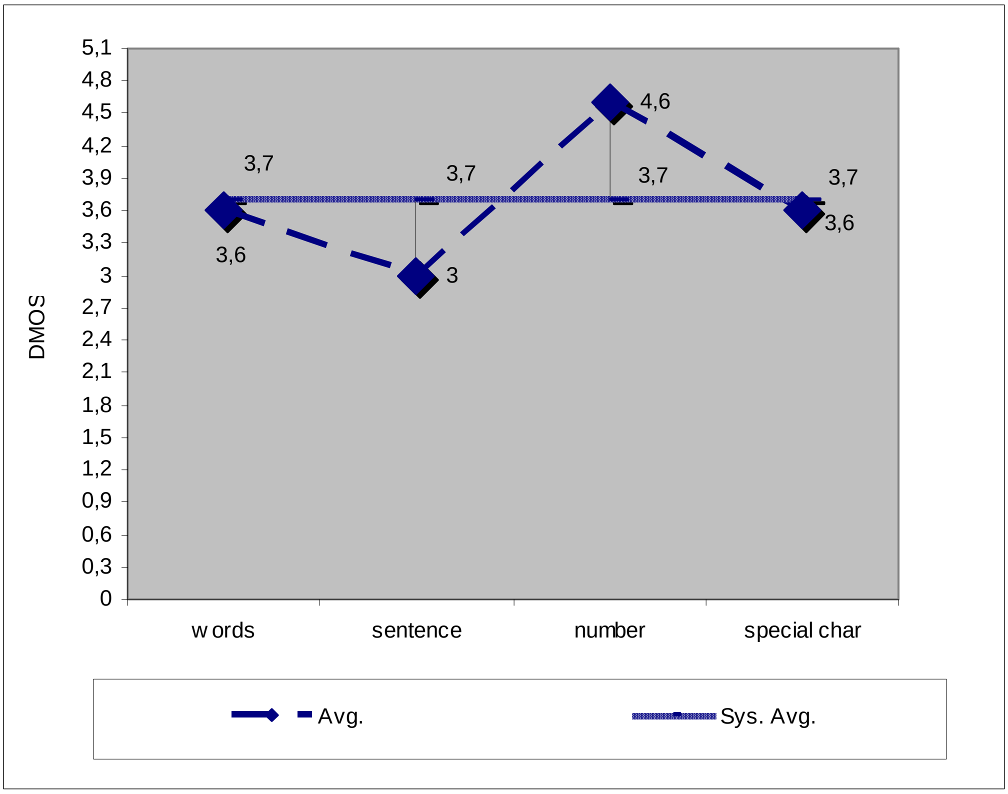


Fig. 4: DMOS scores for all components and the average of the system

5. Conclusion and Future Work

In this study the framework of a TTS system for Turkish is built. Although the system uses simple techniques it provides promising results for Turkish, since the selected approach, namely the concatenative method, is very well suited for Turkish. The system can be improved by improving the quality of the speech files recorded. The sound files of news, films etc can be explored for extracting the recurrent sound units in Turkish instead of recording the diphones one by one. There are some ongoing projects [3] about the analysis of speech signals for various applications and can be helpful for obtaining wide ranges of phonemes in synthesis.

The punctuations are removed in the preprocessing step just to eliminate some inconsistencies and obtain the core system. In the future versions of the TTS, the text can be synthesized in accordance with the punctuations for considering the emotions and intonations as partially achieved in some of the researches [4] . The synthesis of a sentence ending with a question mark can have an interrogative intonation and synthesis of a sentence ending with an exclamation mark can be an amazing intonation. In addition to these, other punctuations can be helpful for approximating the synthesized speech to its human speech form such as pausing at the end of the sentences ending with full stop and also pausing after the punctuation comma.

References:

[1] Aşlıyan R., Günel K., Türkçe Metinler İçin Hece Tabanlı Konuşma Sentezleme Sistemi, *Akademik Bilişim 2008*, Çanakkale Onsekiz Mart Üniversitesi, http://ab.org.tr/ab08/bildiri/116.doc, Retrieved on June 24, 2008

[2] Buza, O., Toderean, C.G., A Romanian Syllable-Based Text-To-Speech System, *Proceedings of the 6th WSEAS International Conference on Signal Processing, Robotics and Automation*, Corfu Island, Greece, February 16-19, 2007, 77-83

[3] El-Imam, Y.A., Elwakil, A.S., Applying recurrence quantification and spectral analysis to represent nasalization in speech signals, *Proceedings of the 7th WSEAS International Conference on Multimedia Systems & Signal Processing*, Hangzhou, China, April 15-17, 2007, 129-133

[4] Kurematsu, M., Hakura, J., Fujita, H., The Framework of the Speech Communication System with Emotion Processing, *Proceedings of the 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases*, Corfu Island, Greece, February 16-19, 2007, 46-52

[5] MOS, http://en.wikipedia.org/wiki/Mean_Opinion_Score, Retrieved on June 24, 2008

[6] Shah, A.A., Ansari, A.W., and Das L., Bi-Lingual Text to Speech Synthesis System for Urdu and Sindhi, *National Conf. on Emerging Technologies*, 2004, http://www.szabist.edu.pk/ncet2004/docs/session%20vi%20paper%20no%204%20(p%20126-130).pdf, Retrieved on June 24, 2008

[7] SIL Software Speech Analyzer, http://www.sil.org/computing/catalog/show_software.asp?id=57, Retrieved on March 10, 2008

[8] Speech Synthesis, http://en.wikipedia.org/wiki/Speech_synthesis, Retrieved on June 24, 2008

[9] ZEMBEREK, http://code.google.com/p/zemberek/, Retrieved on March 1, 2008

[10] Zhang, J., Language Generation and Speech Synthesis in Dialogues for Language Learning, MS thesis, 2004, http://groups.csail.mit.edu/sls/publications/2004/zhang_thesis.pdf , Retrieved on June 24, 2008