A Concatenative Turkish Text-to-Speech System and Evaluation Process

Zeynep Orhan[1], Zeliha Görmez[2]

[1]Fatih University, Computer Engineering Department, Turkey
[2]Fatih University, Vocational School, Turkey
zorhan@fatih.edu.tr, zcetin@fatih.edu.tr

Abstract

In this study, a concatenative text-to-speech system for Turkish is built. The system uses simple techniques and the concatenation units are obtained from the atomic units. This approach is very well suited for Turkish language structure and it is flexible enough to allow the synthesis of all types texts. The Turkish TTS system is tested considering the naturalness and intelligibility criterion. It is evaluated by using MOS, DRT and CT. Naturalness and intelligibility of the Turkish TTS system is tested by MOS and CT-DRT, respectively. Although the system uses simple techniques, it provides promising results for Turkish TTS, since the selected concatenative method is very well suited for Turkish language structure.

1. Introduction

Speech synthesis technology refers to the knowledge of producing the artificial sounds that will be interpreted as speech that can be possibly strange but yet understandable [7]. The human speech can be produced artificially either in software or hardware as the ultimate goal of the speech synthesis. The natural language text is converted into speech by Text-To-Speech (TTS) systems as a sub branch of speech synthesis. Building a system that clearly gets across the message and achieving this by using a human-like voice are the fundamental concerns of a computer system capable of speaking. Within the research community, these goals are referred to as *intelligibility* and *naturalness*, respectively [12].

TTS systems have been widely used as assistive technological tools for a long time. Voice information services (i.e. price lists, weather forecasting reports, etc.), talking PCs, toys, and dictionaries, mobile alert services that speak to warn us for giving information about time and schedule, automated question-answering systems, audible daily journals, books, e-mails that can be used while working, doing housework, traveling, driving, or exercising for saving time are only a few examples of the application areas of TTS that worth mentioning.

The TTS systems are generally composed of various processes. The major processes that take place in a typical TTS system are shown as a block diagram in Fig 1.

The technologies used in TTS systems can basically be categorized into two groups as *formant synthesis* and *concatenative synthesis*[*]. The strengths and weaknesses of both technologies exist depending on the requirements of the systems where they are employed.

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

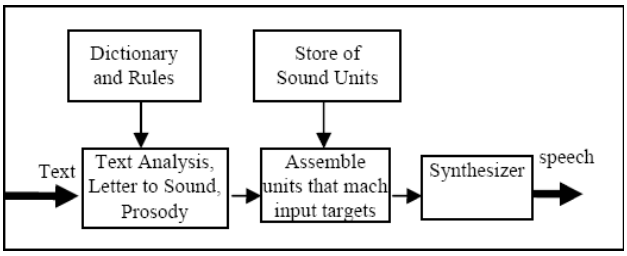

Fig 1. Block Diagram of the TTS [9]

In *formant synthesis* human speech samples are not used at runtime and the vocal tract transfer function is modeled by simulating formant frequencies and formant amplitudes [10]. Formant synthesis provides intelligibility but naturalness is not assured. The memory and microprocessor power requirement is less than the other techniques; therefore it is suitable for the limited devices [11].

On the other hand, the concatenative approach is based on the small pieces of recorded speech. In this approach, to prepare speech database, the small pieces are 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 the 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 and can disturb the intelligibility [11].

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 [14].

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 [11]

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 is chosen for Turkish to keep the size of the database small to have reasonable amount of DSP.

The rest of the paper is organized as follows: The second and the third sections are dedicated to the speech database of Turkish and the specific implementation of the concatenative TTS system. Fourth section explains the details of the evaluation process of the Turkish TTS system. The last section provides concluding remarks and the possible future work for improving the current system.

2. Speech Database

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 syllables, especially the ones that have 5 or 6 characters are very rare. The ratios of the syllables that consist of one-letter is 5.93%, two letters 56.57%, tree-letters 35.16%, four-letters 2.18% and five-letters 0.17%. The percentages of these syllables are obtained from the Turkish corpora prepared in a research of this domain [2]. The most frequently used syllables and their examples are given in Table 1.

Table 1. The structure of Turkish syllables

Syllable structure	Sample syllables
V	a, e, ı, i, o, ö, u, ü
VC	ab, ac, aç, … ,az, eb, ec,…
CV	ba, be, bi,…, za, ze, zı, …
CVC	bak, git, say, kır, …
VCC	ast, üst, ırk, …
CCV	tra, pla, tre
CVCC	Türk, kürt, sırt,

Table 2. Diphones kept in Turkish speech database

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

The speech database of Turkish includes single or double letter sounds as the smallest phoneme in the current system and given in Table 2. The minimum number of phonemes kept in the database is calculated as 365 and the rest of the syllables are formed from the concatenation of these sounds. The rest of the syllables are formed from the concatenation of these sounds. The longer syllables are synthesized as follows:

- Target CVC: CV+C → CV+VC (git/go→gi+it)
- Target VCC: VC+C → VCC (üst/top→üs+t)
- Target CCV: CC+V → C(V)CV (bre‡→ b(i)re)
- Target CVCC: CV+C+C→CV+VC+C(türk→tü+ür+k)

For CVC syllables last consonant affected by previous vowel so VC is used instead of only C leading to a more natural sound.

Speech database contains units that are cut from large speech signals. There are tree forms for each unit because phonemes are also affected by their position in the word. For example the 'şe' syllable:

- First form: şe-ker (sugar) (beginning)
- Middle form : şi-şe-ler (bottles) (middle)
- Last form : kö-şe (corner) (end)

The database consists of alternative diphones for phonemes. The syllable *'ka'* in the word *kağıt(paper)* should be pronounced as *kâ-ğıt* by using a caret or circumflex, but on the other hand *'ka'* in the word *ka-lem(pen)* as it is written.

The speech database contains three versions of 365 diphone records and alternative diphone records for some phonemes in total.

Speech records that constitute the database are obtained by cutting them from the continuous speech of the Turkish native speaker. This continuous speech is constructed manually in an acoustically isolated room is composed of 100 sentences that contain exemplars of all the phonemes and their forms as much as possible.

3. Implementation

There are two major modules in a TTS system: Natural Language Processing Module (NLP) that converts the written text input in phonetic transcription and Digital Signal Processing Module (DSP) that transforms the symbolic information into speech [4]. Turkish TTS system is also formed from these two components that are explained in the following subsections.

3.1. Text-to-Phoneme

The first step of text-to-phoneme is preprocessing. The preprocessing of the text to be converted into speech requires the followings:

- First, all unnecessary format information that does not contribute anything to the pronunciation are cleared and the redundant characters such as extra white spaces (\t, \n, \s) etc. are removed.
- Next, the text is normalized by considering upper/lowercase letters
- Then this step is followed by the syllabification phase. The Zemberek Project [13] is used as a tool for this step in the 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. The pseudocode for the main function of text-to-phoneme phase is shown in Table 3. The pseudocode of sub-functions for syllabification, checking epenthesis and abbreviation function are shown in Table 4, part A, B and C, respectively.

* Consonants in Turkish: b, c, ç, d, f, g, ğ, h, j, k, l, m, n, p, r, s, ş, t, v, y, z

† Vowels in Turkish: a, e, ı, i, u, ü, o, ö

‡ Two consonats can not come one after the other at the beginning of a word in Turkish. However, the words borrowed from other languages via cultural interactions may disturb this and suitable vowel insertion is essential for correct pronounciation.

Table 3. Pseudocode for main function of text-to-phoneme [15]

```
Check spaces, convert extra spaces{\n,\t\r} into single space
  Check liaison, convert 'CspaceV' into CV (delete space between
  C and V)
  Check special character, convert character into
  SpaceCharacterSpace (put character between two spaces)
  Check comma, do nothing if comma between two numbers
  otherwise convert comma into SpaceCommaSpace.
Check spaces again.
Tokenize text, divide text into word array considering space.
 For each token:
if token is a number use as number
else // as a word
  if it has one character
    if the character is a special character convert its vocalization
     else if vowel do nothing
     else if consonant concatenate with 'e' vowel
  else if has two letters
    if it is in the form of CV or VC syllable send zemberekSyllable
   else check epenthesis for foreign word and send zemberekSyllable
```

Table 4. Pseudocode of sub-functions [15]

```
function zemberekSyllable
  declare syllable as text
  convert word into syllable array with zemberek project
  if is syllable array add abbreviation function result to syllable
 for each syllable:
if it has less than three characters add it to syllable
else if it has three characters
  if it is in CVC format add it as CV+C to syllable
  else if CCV format add abbreviation function result to syllable
  else if is VCC format check CCend set
 if is available add it as VC+C to syllable
 else add abbreviation function result to syllable
else if it has four letters(CVCC) check CCend set
  if is available add it as CV+C+C to syllable
  else add abbreviation function result to syllable
```

(A)

```
function epenthesis
 if token matches CC[C|V]+
pattern check CCbegin set
if available insert required
vowel
```

(B)

```
function abbreviation
for each letter:
 if vowel do nothing
 else if consonant add 'e'
vowel to consonant
```

(C)

Table 5. Vocalized and not vocalized special characters

Special characters	
Vocalized	@, %, &, #, *, /, -, +, >, <, (,), =, ~, €, $, _
Not vocalized	., ;, ?, !, ', "

There are some other issues that should be considered during the synthesis of the units. The liaison is the grammatical circumstance in which a usually silent consonant at the end of a word is pronounced together with the vowel at the beginning of the word that follows it. The examples of the liaison are pointed out by → in the following lines of a Turkish poem:

*"Dönülmez**→ak**şamın ufkundayız vakit çok geç*
*Bu son fasıldı**r→ey** ömrüm nasıl geçersen geç"*

In the above lines normal syllabification should result in

Dö-nül-mez ak-şa-mın or *fa-sıl-dır ey*

But due to the liaison they are synthesized as

Dö-nül-me zak-şa-mın or *fa-sıl-dı rey*

However, the following example is not a liaison, since the comma disturbs the rule. Therefore, the punctuations are required for liaison detection. (Liaison is not applicable at → point).

*Anne**m,** **a**blam geldi.*

Special characters that are considered and ignored in synthesis are shown in Table 5. Comma is synthesized if between two numbers as in the case of 11,3 and synthesized as *onbir virgül üç (eleven comma tree),* and ignored if it occurs between words. Exceptionally, the comma is not considered between two numbers when a space follows it. Therefore, 11, 3 is synthesized as two distinct numbers *on bir(eleven) üç(three)*. Not as *onbir virgül üç (eleven comma tree)* as in the previous case.

CCend and *CCbegin* sets shown in Table 6 are possible combinations of two consecutive consonants that can occur at the end and the beginning of a syllable in Turkish. These sets are used for the detection of the abbreviations and epenthesis. If the syllable includes two consecutive consonants at the end, CCend set is checked for the valid combination. If it is valid, it is synthesized as a normal syllable; otherwise it is considered as an abbreviation. The word *fabl (fable)* includes two consonants at the end and synthesized as fa+b+l because 'bl' consonant combination is member of CCend set. On the other hand, *ABS* (Anti-lock Braking System) ending with two consonants which is not a valid combination, since 'bs' consonant combination is not member of CCend set, accepted as an abbreviation and synthesized as *A-Be-Se*.

Table 6. Possible combinations of two consecutive consonants at the end and beginning of a syllable in Turkish

CCend	bd, bl, br, bt, cd, dh, dr, fl, fr, fs, ft, gl, gr, hd, hr, ht, kl, kr, ks, kt, lç, lf, lg, lh, lk, lm, lp, ls, lt, mb, mp, mt, nç, nf, ng, nk, ns, nt, nz, rç, rd, rf, rg, rh, rj, rk, rm, rn, rp, rs, rş, rt, rv, rz, sk, sp, st, şk, şt, tf, tm, tr, vk, vr, vs, vt, yh, yl, yn, yp, yr, ys, yt, zm
CCbegin	br, dr, fl, fr, gl, gr, hr, kl, kr, pl, pr, ps, sf, sk, sl, sm, sp, st, tr

The potential epenthesis is checked especially for foreign words via the CCbegin set. If the syllable includes two consecutive consonants at the beginning, CCbegin set is checked for the valid combination. If it is valid, additional sounds are inserted by using some heuristics and exceptional case rules. Examples are the word *grip(flu),* since *gr* is a member of CCbegin set, the word is converted to *gırip*, by inserting *ı* in between *g* and *r*, or *profesör(professor)* to *purofesör* etc.

The input of this step is a string of Turkish and the output is obtained as arrays of words composed of one or two-letter components of the input that are going to be sent to the second step as the input for synthesis. More specifically if the input is *Bugün hangi gün? (What day is it?),* the output is returned as *[Bu, gü, n] [ha, n, gi] [gü, n]*.

3.2. Phoneme-to-speech

In this step the input is previous step's output as an array such as [*Bu, gü, n] [ha, n, gi] [gü, n]*. These array elements are mapped to their corresponding wav files. Selection of the

appropriate file is achieved depending on the position of the component as in the case of syllable 'şe' for the words şe-ker(sugar), şi-şe-ler(bottles), and kö-şe(corner). If the syllable is C and previous syllable CV (gü, n), C convert VC (gü, ün) because for CVC syllables last consonant affected by previous vowel. The duration of words and silence parts are adjusted and the synthesis process is completed.

4. Evaluation Tests

Several methods have been developed to evaluate the overall quality or acceptability of synthetic speech [6]. Mean Opinion Score (MOS) [5], Comprehension Test (CT), Diagnostic Rhyme Test (DRT) are the most frequently used techniques for the evaluation of the naturalness and the intelligibility of TTS systems. Naturalness and intelligibility of the Turkish TTS system is tested by MOS and CT-DRT respectively.

The MOS is generated by averaging the results of a set of standards, subjective tests where a number of listeners rate the perceived audio quality of test sentences. A listener is required to give each sentence a rating between the range of 1 (Bad) - 5 (Excellent). The perceptual score of the method MOS is calculated by taking the mean of all scores for each sentence [8].

In the context of this study, 16 sentences are used for tests and 27 native Turkish speakers employed in the evaluation. Average MOS values for each sentence are given in Fig 2. The MOS average for the Turkish TTS system is calculated as 3.42.

In the comprehension tests, a subject hears a few sentences or paragraphs and answers the questions about the content of the text, so some of the items may be missed [1]. It is not important to recognize one single phoneme, but the intended meaning. If the meaning of the sentence is understood, the 100% segmental intelligibility is not crucial for text comprehension and sometimes even long sections may be missed [6], [3].

In the comprehension tests three subtests are applied. In all three cases, the testers are allowed to listen to the sentences twice. In the majority of the tests, success is achieved for the first listening trial, and second one also improves the results. Accuracy is calculated as the ratio of number of correct answers given by the testers to the whole set of correct answer. The accuracies of the CT and DRT tests are given in Table 7.

First comprehension subtest has 8 sentences and 8 questions about the content. Listeners answer the question about content of each sentence.

Second subtest is about answering common questions. It contains 8 sentences. Listeners answer the questions. The results indicate that the understandability of the system is very high.

Third subtest is applied as a filling in the blanks test. There are 8 noun phrases, one word of the phrase is provided to the listener and other is left as blank. The testers listened to the speech and filled in the blanks. The results indicate that the system achieved a high understandability rate.

In the DRT test of the current system, the consonants that are similar to each other are selected and the listeners are asked to distinguish the correct consonant among the similar sounding alternatives. The letters that have the same way out such as 'b' and 'p' are plosive and bilabial consonant and can be easily misunderstood.

The similar sounding words that are used for DRT are provided for the listeners and they are asked to choose one word from the given table that is the same with what they hear. The accuracies are calculated above 0.90 and mostly close to 1 especially after the second trial.

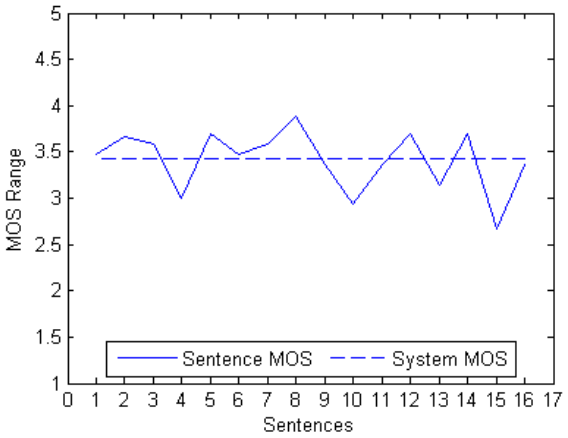

Fig 2. Average MOS values for each sentence and system average

Table 7. Comprehension Tests and DRT accuracy [15]

Tests	**Listening trial accuracies**	
Comprehension Tests	**1st**	**2st**
Answer the questions about the content	0.94	1
Answering common question	1	1
Fillings the blanks	0.97	1
Diagnostic Rhyme Test (DRT)	0.90	1

5. Conclusions

A Turkish TTS system that uses a concatenative synthesis approach is implemented and evaluated in the context of this study. The system uses simple techniques and the concatenation units are obtained from the atomic units. This approach is very well suited for Turkish language structure and it is flexible enough to allow the synthesis of all types texts. The evaluation process that yields high accuracies both for naturalness and intelligibility criterion is carried out by using the MOS, CT and DRT techniques as being the most frequently employed evaluation approaches in this field.

The prosody of continuous speech that is considered melody, rhythm, depends on many separate aspects, such as meaning, speaker characteristic and emotions [6]. The prosodic dependencies are shown in Fig 3. Unfortunately, written text doesn't contain all these features. Sometimes punctuations may be a help for prosody analysis such as exclamation mark or question mark.

The punctuations are removed in the preprocessing step just to eliminate some inconsistencies and obtain the core system. In the future versions they can be used for adding emotions and intonations.

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.

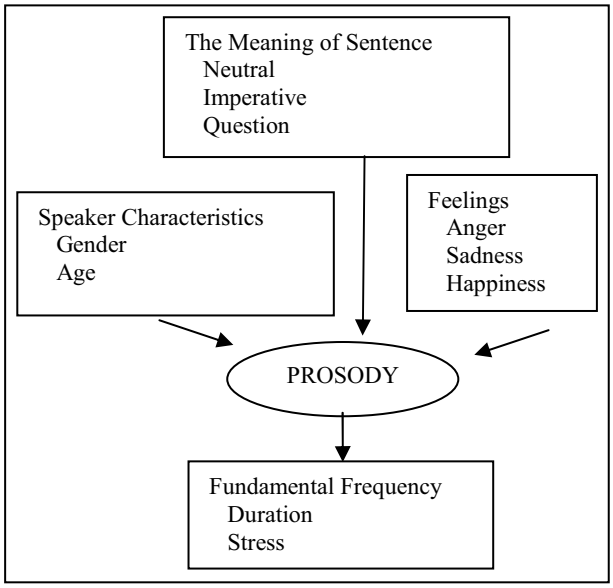

Fig 3. Prosodic Dependencies [6]

6. References

[1] Allen J., Hunnicutt S., Klatt D. (1987). From Text to Speech: The MITalk System. Cambridge University Press, Inc.

[2] Aşlıyan R., Günel K., Türkçe Metinler İçin Hece Tabanlı Konuşma Sentezleme Sistemi, Akademik Bilişim 2008, Çanakkale Onsekiz Mart Üniversitesi,

[3] Bernstein J., Pisoni D. (1980). Unlimited Text-to-Speech System: Description and Evaluation of a Microprocessor Based Device. Proceedings of ICASSP 80 (3): 574-579.

[4] Cerrato, L., "Introduction to speech sythesis", Available*:http://stp.lingfil.uu.se/~matsd/uv/uv05/motis/lc_synt.pdf* , Retrieved on Feb 01, 2009

[5] Goldstein M. (1995). Classification of Methods Used for Assessment of Text-to-Speech Systems According to the Demands Placed on the Listener. Speech Communication vol. 16: 225-244.

[6] Lemmetty S., "Review of Speech Synthesis Technology", MSc. Thesis, Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, Finland, 1999.

[7] Maxey H. D., Smithsonian Speech Synthesis History Project,http://americanhistory.si.edu/archives/speechsynthesis/ss_spsyn.htm

[8] Mean Opinion Score (MOS) terminology. ITU-T P.800.1. 2003/03, Retrieved on 27, 2009

[9] 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

[10] Styger, T., & Keller, E. (1994). Formant synthesis. In E. Keller (ed.), Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges (pp. 109-128). Chichester: John Wiley.

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

[12] Taylor P., Text to Speech Synthesis. University of Cambridge, 2007, DRAFT available: http://mi.eng.cam.ac.uk/pat40/ttsbook_draft_2.pdf, Retrieved on Feb 01, 2009

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

[14] 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

[15] Görmez, Z., "Implementation of a Text-to-Speech System with Machine Learning Algorithms in Turkish", Msc. Thesis, Department of Computer Engineering in Fatih University, 2009