SEGMENTAL INTELLIGIBILITY TEST OF TEXT-TO-SPEECH SERVICES FOR ROMANIAN LANGUAGE BASED ON LOGOTOMS

Olga PUSTOVALOVA

Catedra Tehnologii de Programare

Sinteză vocii oferă posibilitatea de a citi texte în mod automat fără participarea omului. Sisteme de acest tip primesc mesajul de tip text la intrare şi generează semnalul vocal la ieşire. Au fost elaborate diferite aplicaţii de sinteză vocală, folosite acum în citirea paginilor web, cărţilor, sistemelor de navigare GPS pentru automobile etc.

Calitatea vocii generate în mod automat depinde de tehnologiile de creare. Pentru testarea vocilor noi se aplică metode de estimare a calităţii. Teste de tip „*logotoms*” evaluează claritatea segmentală a vocii, anume: claritatea sunetelor separate în diferite poziţii ale cuvântului. Cuvinte de tip „*logotom*” sunt intenţionate de a fi incidente în sens semantic.

În această lucrare este prezentat un test de tip „*logotom*” elaborat pentru limba română.

Introduction

Speech synthesis allows reading a text automatically without involving a human speaker each time. *Text-to-speech* (TTS) conversion systems receive a text on input and generate a *speech waveform* on output, which gives to a computer a possibility to play an audio version of relevant text.

Various applications of speech synthesis have been designed over the years. Speech synthesis is now useful in reading web-pages or e-mail messages, creating audio-books for personal use, listening instructions of GPS navigation systems while driving a car etc. It provides significant help for visually impaired users. Research and development companies work over improvement of speech usage in such areas as call center assistance or e-commerce.

However, the quality of automatically generated speech may vary. In particular, it strongly depends on technologies used for speech generation. Some synthetic voices may sound mechanical, may have unpleasant intonations or unintelligible sounds.

When selecting a voice for a specific use, or when testing a newly developed voice, a number of quality assessment methods could be applied. These methods estimate voice intelligibility or/and voice naturalness. Input data for such measurements is language-dependent.

In this paper, some effective methods of quality assessment are described. Sample data for experiments with Romanian voices are presented.

1. Criteria of quality

Quality measurement needs general agreements: we nee to define, what items are essential for quality measurement and what methods of assessment would be valid. TTS speech quality is usually measured in comparison to performance of another TTS [5], natural voice is also included as TTS. Synthetic voices are usually compared to each other on the following criteria of quality: intelligibility and naturalness [7]. Definitions are represented in their traditional meaning, as presented in sources i.e. [5] and [6].

Intelligibility is capability of synthetic speech of being understood, or comprehended. If speech is not articulated enough, its quality would be low.

Segmental intelligibility is capability to articulate separate sounds clearly. That parameter shows whether speech items which construct words are integrated enough to make these words understandable.

Supra-segmental intelligibility is capability to articulate the whole message clearly, and high intelligibility of a separate sound could be optional in this case.

Naturalness is another measured factor, and it’s usually understood as a way of similarity between human and synthetic speech prosody. It includes intonations, accents, general sounding of speech. Though there are no certain conventions of what naturalness really is ([8], p.2), a number of factors are considered to affect naturalness: occurrence of deviating speech sounds, speaking rate, voice pleasantness, appropriate liveliness etc. ([8], p.5).

Intelligibility and naturalness are both essential for speech quality assessment, and can be measured for estimating both speech presentation and speech perception. Though, other factors may be considered depending on actual problem (i.e. [4]).

2. Materials and methods

2.1. Experiment requirements

Speech quality is subjective; it means that only series of experiments involving human interaction can bring reliable results [9]. A group of selected subjects is meant to statistically represent future users of TTS. Subjects may be asked to repeat pronounced words, write down on a sheet missing parts of words or sentences pronounced by synthetic voice, or evaluate different aspects of TTS performance by filling an opinion questionnaire [9].

The following elements are required for subjective quality assessment of a given TTS:

- **stimuli:** voice samples generated by TTS. Selection of stimuli depends on approach used for quality assessment.
- **participants, or subjects:** people who agreed to participate in session of TTS quality assessment. An important thing is that TTS system’s performance is measured, not subjects’.

In some methods, stimuli must be generated by several different TTS systems (i.e. [5]). Quality of a particular TTS is then defined in reference to other systems’ performance.

2.2. TTS for Romanian language

In quality assessment sessions several TTS can be compared. The following Romanian TTS are available for on-line testing are presented in the Table 1:

Table 1

List of some Romanian text-to-speech engines

Name	Company
IVONA, voice *„Carmen*”	*IVO Software*, Poland http://www.ivosoftware.com/ivonaonline.php
Phobos, based on *MBROLA*	*Phobos Soft*, Romania http://www.phobos.ro/demos/tts/index.html
Baum, voice *„Ancutza*”	*Baum Engineering SRL*, Romania http://www.baum.ro/ro/online/online.html

2.3. Intelligibility tests

Segmental intelligibility tests are designed for measuring intelligibility of separate sounds, and it needs careful preparation of word lists. Supra-segmental intelligibility is used for evaluation of the whole sentences or tests, so that sentences or text preparation is needed [6].

Ojala [6] gives a review of existing intelligibility tests, giving the following classification:

- tests based on meaningful words (phonetically balanced word lists [1]; rhyme tests (i.e.[2]));
- tests based on non-sense words (logotoms) [3].

2.4. Segmental intelligibility test based on non-sense words (logotoms)

Stimuli are presented as template-based words [2] generally having no sense (also named *logotoms*). Word templates are often the following: CVC – for testing consonants in initial and final positions; VCV – for testing consonants in middle position; where *V – vowels, C – consonants*. Other word templates may be used, if necessary [6].

Subjects are asked to repeat words played back to them, or write down missing sounds by filling gaps on a response sheet. The test is semantically unpredictable, and therefore allows assessing genuine sound intelligibility.

3. Experiment

3.1. Experimental model

In use case for Romanian, word set was generated using a table of letter occurrences in Romanian. The task is to disseminate letters aiming to fill available templates (i.e. CVC and VCV) in such a way that their occurrence frequencies [10] would coincide with usual frequencies for the language.

Word templates are coded in input txt-file. Two main symbols may be used: to code a vowel (‘&’) or a consonant (‘#’). All other symbols are not processed and are presented in output file via blank spaces.

Since the task is similar to aleatory variable modeling, we used the built-in Pascal random number generator. Occurrence frequencies and corresponding letters are defined as constant arrays. Generated random number means a point collocated between occurrence frequencies of two alphabet letters, so the letter could be extracted. In fact, values of frequencies are presented as lengths of closed geometrical intervals, and chance of selecting a point which belongs to an interval of a specific letter would be better when letter frequency is larger.

A scheme of realization is given below:

```
repeat goal:=random until match(goal,ltype); {goal is vowel/consonant}
currentPath:=0;
currentLetter:=1;

while currentPath < goal do
begin
currentPath:= currentPath + space[currentLetter];
currentLetter:=currentLetter+1;
end;

getLetter:= letter[currentLetter-1];
```

Here *letter[]* is an array of letter characters, *space[]* is an array of corresponding letter frequencies. Data contained in constant arrays are optimized so that letters with higher frequencies are positioned earlier.

Each generated random letter is written in output file according to input file content.

Any number of intelligibility tests can be generated by such a mechanism.

3.2. Obtained data set

Word set was balanced to provide standard occurrences for each vowel and consonant, both in initial and final position. Obtained example results are presented in Table 2 and Table 3.

Table 2

Segmental intelligibility test for Romanian – logotoms for VCV template

uvi	âle	îce	ici	Alu	idi	ibă	Eci	ele	alu
ane	eto	ade	ure	Ule	iri	ici	Ăci	uri	iră
ede	ada	ade	eră	Ina	uti	ate	Ute	eta	esu
odi	efa	aru	emo	Ute	ară	eci	Ula	înî	eru
ute	ilu	ufu	eru	emu	alo	oma	Ure	eli	ucâ
ubi	ili	ălo	olu	ără	eşi	ane	Uvi	ecă	epe
ara	ore	emo	ore	ără	ătâ	eră	Uta	ati	ini
alu	ito	ana	edâ	ali	eca	ate	Epa	ăna	îce
eci	etu	ană	ipa	ere	idi	isu	Ode	upa	isi
ate	ivi	uri	ato	ecă	efi	âma	Uni	ari	use
ilu	egă	ato	oda	ato	opi	ata	Ema	ămi	era
işi	adâ	eta	aşî	ito	ăre	ena	Uli	ăne	ena
uțe	iro	oră	ena	ele	aci	ifă	Uda	ato	ele
ono	ele	ită	ate	île	ita	aca	Eco	ibi	ato
ire	eja	ăta	ici	idi	ăsu	ite	Ăre	ita	ire
ece	âci	ace	oso	ari	ada	iza	Ubă	ătî	uge
era	îpe	ige	ită	ală	usu	ina	Eta	ulo	iri
udi	eti	ăco	ută	adu	ase	ăpi	Ine	yni	ati
ime	ili	ăle	imă	ule	esa	ene	Ădo	ofe	ivu
ome	eni	elă	asa	uda	eto	evi	Ara	itu	işi
avă	ure	eli	idi	itu	isu	ape	Opa	iva	ăru
ăgă	ună	ura	ore	âgu	ăse	utu	Aco	una	ili
odi	ada	eva	ărî	ifi	acă	ula	Asa	ana	uto
opi	eta	ici	esi	oco	awu	obe	Uni	eca	aci
îlî	enâ	ase	ăre	ilu	una	ise	Ise	amo	ofi

Table 3

Segmental intelligibility test for Romanian – logotoms for CVC template

măn	*pîț*	*sel*	*rit*	*nic*	*tim*	*tip*	*Loh*	*sad*	*cuv*
ris	*tăl*	*tev*	*tan*	*dil*	*nad*	*cit*	*Nas*	*tîn*	*til*
rac	*nin*	*sad*	*per*	*tiş*	*rel*	*tag*	*Lâc*	*sal*	*păt*
nif	*mâl*	*ter*	*rid*	*cig*	*vip*	*per*	*Cec*	*şab*	*xal*
tip	*cit*	*vin*	*puf*	*ses*	*vez*	*lec*	*Şan*	*sos*	*ces*
sun	*tac*	*dem*	*ris*	*cas*	*luv*	*men*	*Sud*	*cum*	*tac*
lic	*car*	*nip*	*rod*	*şun*	*năl*	*tup*	*Ref*	*lir*	*tar*
sil	*rid*	*dan*	*cen*	*daz*	*lad*	*xan*	*Gin*	*mer*	*tap*
şam	*din*	*gan*	*hev*	*răl*	*râg*	*lir*	*Pun*	*dil*	*mâl*
neț	*car*	*mec*	*pul*	*lig*	*pâş*	*lan*	*Lev*	*doş*	*tut*
rel	*tis*	*nes*	*nor*	*păş*	*lut*	*rez*	*Sel*	*ten*	*şeț*
del	*fef*	*nac*	*ded*	*lâr*	*răs*	*cer*	*Neb*	*lal*	*săc*
ted	*sit*	*năl*	*nat*	*cal*	*şăr*	*tîc*	*răr*	*tir*	*tap*
rac	*nin*	*cal*	*mal*	*zif*	*pip*	*lam*	*măm*	*las*	*tuj*
lab	*ruc*	*pen*	*bon*	*tas*	*met*	*sîs*	*căs*	*cil*	*tat*
toc	*mal*	*suc*	*mav*	*pet*	*sav*	*pen*	*cin*	*let*	*cop*
min	*tam*	*dăf*	*nuf*	*cuş*	*nit*	*tav*	*man*	*rut*	*gîn*
rez	*toc*	*cis*	*lel*	*tev*	*men*	*tiv*	*fap*	*dal*	*sos*
dul	*tat*	*lec*	*şad*	*nat*	*mec*	*tir*	*ter*	*moş*	*pec*
păz	*păd*	*rer*	*col*	*des*	*cin*	*nir*	*şun*	*țar*	*paş*
rel	*cut*	*mec*	*ten*	*tet*	*tud*	*ril*	*nas*	*şer*	*rîs*
lun	*tuc*	*nir*	*car*	*put*	*reş*	*pan*	*tîr*	*pos*	*leg*
pîp	*mav*	*rez*	*nid*	*bit*	*jax*	*nin*	*pot*	*xad*	*het*
măț	*piv*	*vug*	*tup*	*lit*	*şiv*	*răl*	*bir*	*lis*	*tus*
ler	*mip*	*măt*	*mas*	*şăl*	*țaf*	*cal*	*răr*	*căr*	*teş*

4. Discussion

Generated test can be evaluated by accuracy of occurrence frequencies. Dimension of evaluated letter set would be 1500 letters (in words presented in Table 2 and Table 3).

Table 4 presents the following information:

Column 1. Letter.

Column 2. Percentage of its occurrence in the language, cited from [10].

Column 3. Occurrences of the letter in the evaluated word set (table 2 and table 3).

Column 4. Percentage of the letter occurrence in the evaluated word set.

Column 5. Difference between 2 and 4. Easy to notice, that maximum difference values are 1,45 (‘L’), 1,09 (‘T’) and 1,07 (‘N’) - all for consonants.

Table 4

Results of data set evaluation

#	1	2	3	4	5	#	1	2	3	4	5
1	E	11.47	171	11,40	0,07	17	Î	1.40	19	1,27	0,13
2	I	9.96	164	10,93	-0,97	18	V	1.23	25	1,67	-0,44
3	A	9.95	161	10,73	-0,78	19	F	1.18	17	1,13	0,05
4	R	6.82	91	6,07	0,75	20	B	1.07	11	0,73	0,34
5	N	6.47	81	5,40	1,07	21	Ţ	1.00	7	0,47	0,53
6	U	6.20	93	6,20	0,00	22	G	0.99	14	0,93	0,06
7	T	6.04	107	7,13	-1,09	23	Â	0.91	15	1,00	-0,09
8	C	5.28	74	4,93	0,35	24	Z	0.71	8	0,53	0,18

#	1	2	3	4	5
9	L	4.48	89	5,93	-1,45
10	S	4.40	56	3,73	0,67
11	O	4.07	58	3,87	0,20
12	Ă	4.06	68	4,53	-0,47
13	D	3.45	49	3,27	0,18
14	P	3.18	46	3,07	0,11
15	M	3.10	40	2,67	0,43
16	Ş	1.55	24	1,60	-0,05
25	H	0.47	3	0,20	0,27
26	J	0.24	3	0,20	0,04
27	X	0.11	4	0,27	-0,16
28	K	0.11	0	0,00	0,11
29	Y	0.07	1	0,07	0,00
30	W	0.03	1	0,07	-0,04
31	Q	0.00	0	0,00	0,00

Conclusion

Depending on assessment goals, different methods can be used. Logotom tests can help in assessment of TTS segmental intelligibility. *Segmental* intelligibility is measured, it means that attention is focused on intelligibility of sounds occurred within a word, be it initial, middle or final position of sound in a word. Quality of consonants in initial and final positions is the most important. Words are semantically unpredictable and therefore a listener is not likely to guess unintelligible sounds. Naturalness is not measured.

A logotom test model for Romanian language was presented. It statistically represents letter occurrences in Romanian, so that each letter would be tested according to its frequency in the language.

Repherences:

1. Egan J. Articulation testing methods // Laryngoscope. - 1948. - No58. - P.955-991.
2. Fairbanks G. Test of phonemic differentiation: The rhyme test // Journal of the Acoustical Society of America. - 1958. - No30(7).
3. Goldstein M. Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener // Speech communication. - 1995. - No16. - P.225-244.
4. Hartikainen M., Salonen E.P., Turunen M. Subjective Evaluation of Spoken Dialogue Systems Using SERVQUAL Method. - In: Proceedings of ICSLP, 2004, p.2273-2276.
5. ITU-T Recommendation. Telephone transmission quality subjective opinion tests. A method for subjective performance assessment of the quality of speech voice output devices, 1994, p.85.
6. Ojala T. Auditory Quality Evaluation of Present Finnish Text-to-Speech Systems. Sähkö – ja tietoliikennetekniikka // Electrical and Communications Engineering. June, 2006.
7. Schroeter J. The Fundamentals of Text-to-Speech Synthesis // IEEE-ISTO Voice XML Review. - Vol.1. - Iss.3. - 2001.
8. Tatham M., Morton K. Developments in Speech Synthesis. - Wiley, 2005. - 356 p.
9. Viswanathan M., and Viswanathan M. Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale // Computer, Speech and Language. - Vol.19. - 2005. - P.55-83.
10. Vlad A., Mitrea A., and Mitrea M. Two frequency-rank laws for letters in printed Romanian // Procesamiento del Lenguaje Natural. - 2000. - No26. - P.147-153.

Prezentat la 28.03.2008