Evaluation of Articulatory Speech Synthesis: A Perception Study

Dominik Bauer, Peter Birkholz, Jim Kannampuzha, Bernd J. Kröger

Department of Phoniatrics,Pedaudiology and Communication Disorders, University Hospital Aachen and RWTH Aachen University, Email: {dobauer, pbirkholz, bkroeger}@ukaachen.de

Introduction

The method of auditory evaluation by a human listener is often used to check the intelligibility and naturalness of synthesized speech. In our perception study we tested the intelligibility of monosyllabic utterances, generated by an articulatory speech synthesizer. The results are currently used to improve the synthesizer system, which is still under development.

The stimuli were syllables instead of words and sentences to avoid the influence of redundancy of connected speech. No morphological or syntactic information can be used to identify a stimulus. It is important to have a human listener, who is not directly involved in the development of synthetic speech, since there is a strong adaption to the synthesized speech. The operator normally performs significantly better in the recognition task than an untrained listener.

Method

The synthetic monosyllabic stimuli were generated by the articulatory synthesizer software "Speak" (Birkholz 2005, Birkholz et al. 2007, Kröger and Birkholz 2007) using a resynthesis method (Bauer et al. 2009). Syllables produced by a human speaker served as reference in terms of segment duration, accent and intonation.

The set of stimuli contained CV syllables with all voiced and voiceless plosives, nasals and a lateral (/b, d, g/, /p, t, k/, /m, n/, /l/) combined with all qualities of long vowels in Standard German (/i/, /e/, /a/, /o/, /u/) (→ 45 items) and CCV syllables with plosives as the first consonant and the lateral as the second consonant (/bl/, /gl/, /pl/, /kl/) in 5 vowel contexts (→ 20 items). The CCV syllables matched the phonotactic constraints of Standard German. For this reason the corpus does not contain syllables like */dli/ or */tla/.

The stimuli were presented in random order to a human listener who had to spell the syllable he understood. The test was repeated three times on three successive days. To compare the results with the recognition of natural utterances, we also evaluated the intelligibility of syllables produced by a human speaker.

Results

The recognition of natural utterances reached nearly 100%. There was only one confusion for the syllable [mi] which was heard as [ni]. Recognition of synthetic syllables was significantly worse. (See Table 1 for confusion matrix.)

In the group of voiced plosives the recognition rate was 67% for [b], 40% for [d] and 53% for [g]. The recognition of the vowel was not observed at this point.

int	r1	r2	r3
ba	ba	ba	ba
be	be	be	be
bi	bi	bi	bi
bo	vo	vo	vo
bu	u	bu	u
da	da	da	da
de	be	de	be
di	bi	bi	bi
do	o	o	o
du	du	du	du
ga	ba	ba	ba
ge	ge	e	ge
gi	e	i	gi
go	go	go	go
gu	gu	gu	u
pa	pa	pa	pa
pe	pe	pe	pe
pi	ti	pi	ti
po	po	po	wo
pu	ku	ku	ku
ta	pa	pa	pa
te	te	te	ti
ti	ki	pi	ki
to	po	po	po
tu	ku	ku	ku
ka	ka	ka	ka
ke	te	ke	ke
ki	ti	ki	ti
ko	ko	ko	ko
ku	ku	ku	ku
ma	ma	ma	ma
me	me	me	me
mi	mi	mi	mi

int	r1	r2	r3
mo	mo	mo	mo
mu	mu	hnu	mu
na	na	na	na
ne	ne	ne	ne
ni	ni	ni	ni
no	no	no	no
nu	hnu	nu	nu
la	la	la	la
le	le	le	le
li	li	li	li
lo	lo	lo	lo
lu	lu	lu	lu
bla	bla	bla	bla
ble	ble	ble	ble
bli	bli	bli	bli
blo	blo	blo	glo
blu	blu	blu	blu
gla	la	la	la
gle	gli	dle	ble
gli	bli	bli	bli
glo	dlo	glo	glo
glu	lu	glu	blu
pla	ksa	ksa	ksa
ple	tse	kle	kle
pli	ksi	ksi	kli
plo	klo	klo	klo
plu	klu	klu	klu
kla	kla	kla	kla
kle	tle	kle	kle
kli	sli	kli	kli
klo	klo	klo	klo
klu	klu	klu	klu

Table 1: Table of results. int: intended syllable, r1-r3: response in different trials

Recognition rate was 60% for [p], 20% for [t] and 80% for [k]. Nasals and Laterals were recognised significantly better.

[m] reached 93% and [n] as well as [l] reached 100% correct identification.

The plosive [b] in the complex syllables with lateral was identified correctly in 93% of the cases, [k] 87%, [g] 26% and [p] 0%.

Table 2 shows the intended segments or clusters and the number of certain identifications. Correct identifications are marked with grey boxes.

intention \ percept	b	d	g	m	n	l	p	t	k	bl	gl	pl	kl	other
b	10													5
d	6	6												3
g	3		8											4
m				14	1									
n					15									
l						15								
p							3	2	3					1
t							7	3	5					
k								3	12					
bl										15				
gl										5	4			6
pl												0	9	6
kl													13	2

Table 2: intended vs. percepted segments/clusters

Discussion

The perception of a consonant, especially of a voiceless plosive, mainly depends on the acoustic transition phase towards the following vowel produced by the movement of articulators, changing the acoustic properties of the vocal tract [5]. After comparison of the synthesized spectrogram with the spectrograms of a natural signal we expect an improvement on the acoustic level by reducing the bandwidth of the formants in the transitions, increasing the intensity of the plosive burst and increasing the intensity of aspiration for voiceless plosives.

The bandwidth of the formants in the transitions seemed to be too broad due to a wall stiffness constant in the acoustical simulation that is not in account with the values in reality. The values were taken from [5]. Ishizaka et al. measured the stiffness constant at the inner side of the cheek. Values in the synthesizer should be changed with respect to the higher stiffness of the hard palate and the teeth. With these changes, the bandwidth of the formants in the transitions will be reduced.

Another important point to improve the synthesis is to increase the articulator velocity during the release of a plosive. If the velocity is too low, the plosive burst is very weak because of the slowly changing pressure. A way to an improved articulator velocity in plosive releases may be a changed control model, which incorporates the idea of virtual targets.

The confusion of fortis and lenis plosives also indicates that the aspiration parameters in the acoustic simulation should be changed. Especially the amplitude of friction in voiceless plosives should be increased.

Acknowledgements

This work was supported in part by the German Research Council DFG grant Kr 1439/13-1 and grant Kr 1439/15-1.

References

[1] Bauer, D., Kannampuzha, J., Kröger, B.J.: Articulatory Speech Re-Synthesis: Profiting from Natural Acoustic Speech Data. In: A. Esposito and R. Vích (Eds.): Cross-Modal Analysis, LNAI 5641, Springer-Verlag, Berlin Heidelberg (2009) 344–355

[2] Birkholz, P., Kröger, B.J.: Vocal Tract Model Adaptation Using Magnetic Resonance Imaging. Proceedings of the 7th International Seminar on Speech Production. Belo Horizonte, Brazil (2006) 493-500

[3] Kröger, B.J., Birkholz, P.: A Gesture-Based Concept for Speech Movement Control in Articulatory Speech Synthesis. In: Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M. (eds.): Verbal and Nonverbal Communication Behaviours. LNAI 4775, Springer Verlag, Berlin Heidelberg New York (2007) 174-189

[4] Kröger, B.J., Schröder, G., Opgen-Rhein, C.: A Gesture-Based Dynamic Model Describing Articulatory Movement Data. J. Acoust. Soc. Am. 98 (4) (1995) 1878-1889

[5] Strange, W.: Dynamic specification of coarticulated vowels spoken in sentence context. Journal of the Acoustical Society of America, 91, (1989) 2135-2153

[6] Ishizaka, K., French, J.C., Flanagan, J.L.: Direct Determination of Vocal Tract Wall Impedance. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-23, No. 4, August 1975