

Towards Audiovisual TTS in Estonian

Einar MEISTER ^{a,1}, Sascha FAGEL ^b and Rainer METSVAHI ^a

^a*Institute of Cybernetics at Tallinn University of Technology, Estonia*

^b*zoobe message entertainment GmbH, Berlin, Germany*

Abstract. In the current paper we report our first results in the development of audiovisual speech synthesis for Estonian. The MASSY model, developed originally for German, serves as a prototype for the Estonian AV synthesis. First, we give an overview of the methods of AV speech synthesis and the Estonian viseme inventory, then we introduce the MASSY model and its adaptation for Estonian; finally, we discuss the ideas for further development.

Keywords. speech synthesis, talking head, articulation, viseme, Estonian

Introduction

The nature of human speech is essentially bimodal involving auditory and visual information. In face-to-face communication we see and hear the interlocutor and in addition to message exchange we are able to perceive the emotional state and reactions of the partner. Research has shown that in the presence of visual cues, such as lip and tongue movements, speech intelligibility is improved, especially in a noisy environment or in the case of non-native speakers [1], [2], [3], [4], [5]. However, the contribution of visual information in the perception of different sounds is variable. The visual information is most important in distinguishing labiodentals and bilabials (e.g. /t/ and /p/) [6], but e.g. the bilabials /p/ and /m/ cannot be distinguished visually. An example of the complementary nature of multimodal speech perception is the McGurk illusion [7] in which an audio /ba/ paired with a video /ga/ is heard as /da/.

Audiovisual text-to-speech synthesis (AVTTS) is a technique for automatic generation of voice-synchronised facial animations from an arbitrary text. AVTTS has been a topic of active research for several decades and has resulted in applications such as virtual talking heads, animated tutors, aids for the hearing impaired, multimodal interfaces, and others (see e.g. <http://www.speech.kth.se/multimodal/>). The list of languages involved include English, German, French, Italian, Spanish, Swedish, Mandarin Chinese, Japanese, Danish, Czech, etc. (see e.g. <http://mambo.ucsc.edu/psl/international.html>); previously no work on AVTTS has been done for Estonian.

In this paper we introduce the first results in the development of audiovisual speech synthesis for Estonian exploiting a parametric head model.

¹Corresponding Author: Einar Meister; E-mail: einar@ioc.ee.

1. Methods of Audiovisual Speech Synthesis

AVTTS involves two main components: (1) a text-to-speech module to produce synthetic audio of the input text, and (2) a face model producing visible articulatory movements synchronized with audio. Two main approaches have been used for the visual component of AVTTS: (1) a parametric or model-based approach exploiting 2D or 3D graphical facial model, and (2) an image-based approach based on selection of a sequence of video images.

One of the earliest **parametric models** was developed by Parke [8]. The model was implemented as a limited number of meshes representing the face topology and had a small set of parameters to control the lips, teeth, and jaw during speech-synchronized animation. There are several descendants of Parke's model with more advanced design and with a larger number of control parameters, such as BALDI [9], MASSY [10], LUCIA [11], and others, e.g. models developed at KTH [12]. The parametric models do not aim to model the physiological articulatory processes, but only attempt to reproduce the movements of visible articulators synchronously with synthetic audio generated by a separate text-to-speech module. Typically, the control parameters of these models are acquired from the articulatory motions of human speakers by measuring movements of relevant facial points from video recordings, by tracking the markers attached to a speaker's face with a motion capture system, or by using more sophisticated techniques, such as electropalatography, magnetic resonance imaging, ultrasonography, and electromagnetic articulography.

Several parametric models are compatible with the industrial MPEG-4 standard [13] that has defined 84 Feature Points (FPs) to enable the animation of a face model. The FPs are controlled by a set of Facial Animation Parameters (FAPs) which represent a complete set of basic facial actions including head motion, tongue, eye and mouth control [13: 20]. However, the FAPs do not take into account all speech-specific gestures and thus do not provide enough freedom to control the lips and the jaw [14].

The **image-based approach** exploits a large set of annotated video recordings and concatenates single images or video sequences to produce the appropriate visual output. This approach is implemented in e.g. [15], [16], [17].

A more recent approach is **corpus-based AV synthesis** which extends the unit-selection strategy developed initially for audio speech synthesis [18] to audiovisual speech. This method needs a large appropriately annotated bimodal speech corpus in which the acoustic and visual components are kept together. The system searches in the corpus for the most suitable sequence of audiovisual segments that match the target speech. This results in maximal coherence between the two components of bimodal speech and avoids perceptual ambiguity. This approach is introduced in e.g. [19], [20], [21].

In addition, there are attempts to merge the parametric and corpus-based methods resulting in a photo-realistic model with the ability to control the articulatory features, such as lips, tongue and glottis/velum [22].

2. Estonian Phoneme and Viseme Inventories

Estonian has 9 vowel and 17 consonant phonemes [23]. The description of articulatory features of Estonian phonemes is presented in Table 1. All vowels and consonants can occur in the foot structures representing quantity oppositions (vowels in the primary stressed syllable and consonants in the intervocalic position). However, the basic articulatory patterns of phonemes in different quantities remain the same.

Table 1. Articulatory description of Estonian phonemes

Phoneme		Description
Estonian phonological transcription	SAMPA transcription	
/i/	i	illabial high front vowel
/e/	e	illabial mid-high front vowel
/a/	{	illabial low front vowel
/ü/	y	labial high front vowel
/õ/	2	labial mid-high front vowel
/u/	u	labial high back vowel
/o/	o	labial mid-high back vowel
/õ/	7	illabial mid-high back vowel
/a/	A	labial low back vowel
/p/	p	bilabial voiceless plosive
/t/	t	denti-alveolar voiceless plosive
/tʰ/	tʰ	denti-alveolar voiceless palatalised plosive
/k/	k	palato-velar voiceless plosive
/f/	f	labiodental voiceless fricative
/v/	v	labiodental voiced fricative
/s/	s	alveolar voiceless fricative
/sʰ/	s	alveolar voiceless palatalised fricative
/ʃ/	S	postalveolar labialised voiceless fricative
/h/	h	glottal-oral voiceless fricative
/m/	m	bilabial voiced nasal
/n/ *	n	alveolar voiced nasal
/nʰ/	nʰ	alveolar voiced palatalised nasal
/l/	l	alveolar-postalveolar voiced lateral
/lʰ/	lʰ	alveolar-postalveolar voiced palatalised lateral
/r/	r	alveolar voiced trill
/j/	j	palatal approximant

* /n/ has a context-dependent allophone pronounced as [ŋ] before /k/, in SAMPA transcription marked as N

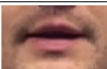
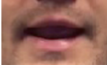
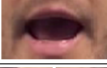









The visually (nearly) indistinguishable phones are grouped to visemes that represent the configuration of articulators corresponding to one or more phones.

The preliminary phoneme-to-viseme mapping was carried out using video recordings of a native male speaker. The speaker read a list of two-syllable CVCV words containing CVC and VCV combinations of Estonian phonemes. All CVCV structures were read in three quantity oppositions in two sets: (1) quantity distinction carried by V1 resulting in structures CVCV (Q1), CVVCV (Q2), CVV:CV (Q3), and (2) quantity distinction carried by the intervocalic consonant resulting in structures CVCV (Q1), CVCCV (Q2), CVC:CV (Q3). Recordings were carried out in a sound-proof room using a Sony HDR-SR12E video recorder.

The recordings were analysed in ELAN software² – a professional tool for the creation of complex annotations on video and audio resources [24]. For each phoneme a frame representing the typical visual pattern of the mouth and the lips was extracted for further measurements. In the case of vowels the typical frame was selected around the midpoint of the primary stressed vowel of Q3 structures, in the case of consonants around the midpoint of the intervocalic consonant of Q3 structures. These selection criteria were motivated by the fact that the duration of these segments is the longest among different quantity oppositions and the quality of a sound around the midpoint should not be affected by the neighbouring sounds.

The initial phoneme-to-viseme mapping by visual comparison and inspection of the characteristic mouth and lip patterns suggests that 26 Estonian phonemes can be represented by 12 basic visemes.

Table 2. The visemes corresponding to Estonian phonemes.

Viseme number	Phonemes in SAMPA	Typical mouth / lip patterns
1	i, j	
2	e	
3	{	
4	u, y	
5	o, ɔ	
6	ɤ, k	
7	A, h	
8	t, t', s, s', n, n'	
9	r, l, l'	
10	m, p	
11	v, f	
12	ʃ	

²<http://www.lat-mpi.eu/tools/elan/>

3. Prototype of the Estonian AVTTS

Parametric talking head models can be easily extended to new languages. For example, Baldi speaks in addition to English also Spanish, Italian, French, German, Mandarin, and Arabic [25]; MASSY, originally developed for German, has been adapted to English, as well [26]. The MASSY model [10] was chosen to serve as a basic prototype for Estonian AVTTS due to its simple and modular architecture.

The basic system consists of four modules: (1) phonetic articulation, (2) audio synthesis, (3) visual articulation, and (4) the virtual face (Figure 1). For the adaptation of the system to Estonian the first three modules must be replaced by the language-specific modules, the face model is language-independent and does not need any adjustment.

As the phonetic module the Estonian txt2pho module developed for the Estonian diphone-based TTS system [27] is used. It generates from the input text the corresponding phone sequence (in SAMPA transcription) together with duration of phones and pauses, and pitch information. The audio signal is synthesized by the MBROLA speech synthesizer [28] exploiting the Estonian diphone database.

The main focus of adaptation is on the visual articulation module which generates the language-specific control parameters for the virtual face.

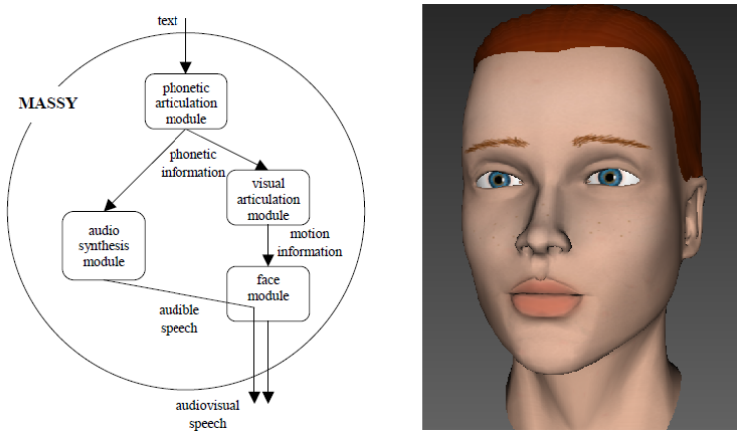


Figure 1. Left: schematic system overview of the Modular Audiovisual Speech SYnthesizer MASSY, right: front view of the MASSY model.

3.1. Face Model

MASSY's 3D face model is one of the descendants of Parke's model [8]. To look as natural as possible the articulatory motion data from a human speaker captured by the Carstens AG 100 electromagnetic articulograph³ (EMA) has been used in its development. From EMA data the motion patterns of the six virtual articulators implemented in the model are derived. The face model produces the articulatory movements according to the control parameters provided by the visual articulation model and adds the audio signal from the speech synthesiser to create the coherent audiovisual speech output. The 3D face model is implemented in VRML (Virtual Reality Modeling Language) [29].

³Carstens Medizinelektronik, <http://www.articulograph.de/>

3.2. Visual Articulation Module

The visual articulation module gets its input from the phonetic module and calculates the motion parameters for the articulators of the virtual head. The six motion parameters of the face model are: **lip width, jaw height, lip height, tongue tip height, tongue back height, and lower lip retraction.**

The motion parameters are generated with a simplified **dominance model** adapted from [30]. It calculates for each viseme the real target position of each articulator based on a fictitious ideal position of this articulator and the strength (the dominance) to control it. The ideal target positions of all articulators for each viseme are language-specific and in the case of the German model they were determined from EMA measurements and video recordings (for details see [10]).

In the first step, the dominance model for Estonian was derived from the German model by adjusting the target values of viseme parameters using the measures of lip spreading and mouth opening calculated from the video frames represented in Table 2. The further tuning of other parameters was carried out in a series of live experiments involving different vowel and consonant combinations, real words and sentences synthesised with the current prototype. The target values of the Estonian viseme parameters constitute an essential part of the current Estonian dominance model.

Clustering of Estonian phonemes/visemes on the basis of articulatory features determined in the current dominance model is presented in Figure 2.

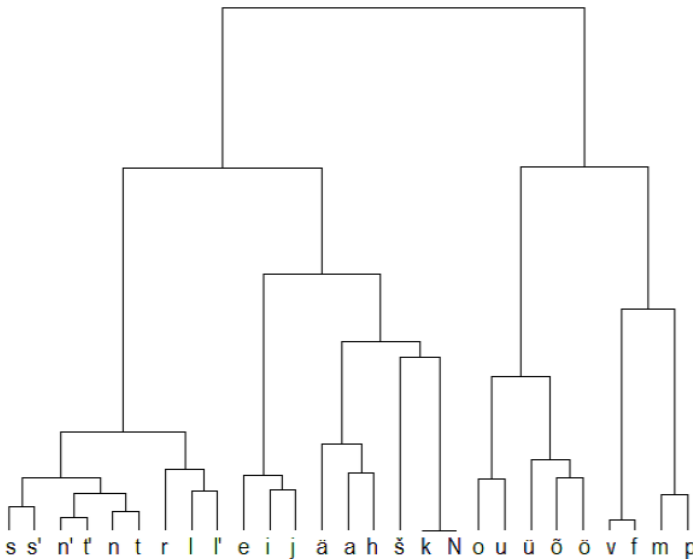


Figure 2. Cluster analysis of Estonian phonemes on the basis of articulatory features.

Figure 3 illustrates the articulatory patterns generated with the current Estonian MASSY model compared to the face patterns of the human speaker in the production of Estonian vowels.

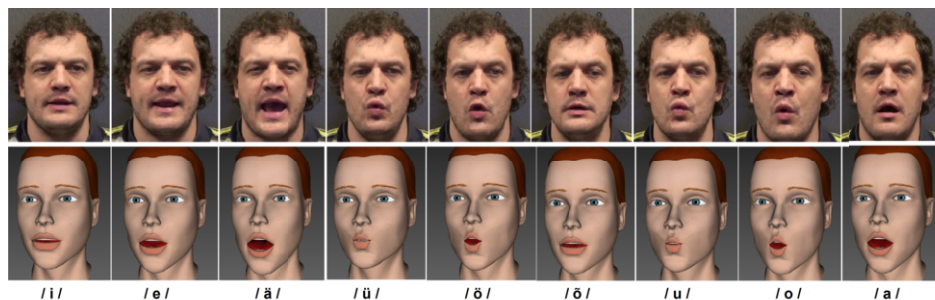


Figure 3. The face patterns of the human speaker and the MASSY model in the production of Estonian vowels.

4. Further Developments

The current prototype of the Estonian AV synthesis is a laboratory prototype and not yet mature for wider use, it will rather serve as the baseline system for further developments. The improvement of the system is possible basically in two ways. First, by the development of a more advanced dominance model; it would need more detailed articulatory data (EMA, electropalatography) which will be available in the multimodal corpus of Estonian speech production [31]. The second alternative would be to integrate a more advanced audio-only text-to-speech system for Estonian, e.g. the system based on unit-selection [32].

Acknowledgements

This work has been supported by the National Program for Estonian Language Technology and by the target-financed theme No. 0140007s12 of the Estonian Ministry of Education and Research.

References

- [1] N. Erber. Interaction of Audition and Vision in the Recognition of Oral Speech Stimuli. *Journal of Speech and Hearing Research* **12** (1969), 423–425.
- [2] C. Benoit, T. Mohamadi, and S. Kandell. Effects of Phonetics Context on Audio-Visual Intelligibility. *Journal of Speech & Hearing Research*, Vol **37** (1994), 1195–1203.
- [3] D. W. Massaro. *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press, 1998.
- [4] S. Fagel, G. Bailly, and F. Elisei. Intelligibility of Natural and 3D-Cloned German Speech. *Proceedings of the AVSP-2007*.
- [5] M. Fitzpatrick, J. Kim, and C. Davis. The effect of seeing the interlocutor on auditory and visual speech production in noise, *Proceedings of the AVSP-2011*, 31–35.
- [6] J. Beskow, M. Dahlquist, B. Granström, M. Lundeberg, K-E. Spens, and T. Öhman, T. (1997). The Teleface project – Multimodal speech communication for the hearing impaired. In Kokkinakis, G., Fakotakis, N., and Dermatas, E. (Eds.) *Proceedings of Eurospeech'97*, (1997), 2003–2006.
- [7] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature* **264** (1976), 746–748.
- [8] F. I. Parke. Parametrized models for facial animation. *IEEE Computer Graphics*, **2(9)**, (1982), 61–68.
- [9] D. W. Massaro. *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, Massachusetts: MIT Press, 1998.

- [10] S. Fagel and C. Clemens. An Articulation Model for Audiovisual Speech Synthesis – Determination, Adjustment, Evaluation. *Speech Communication* **44** (2004), 141–154.
- [11] P. Cosi, A. Fusaro, and G. Tisato. LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaros Labial Coarticulation Model, *Proceedings of Eurospeech 2003*, Geneva, Switzerland, Vol. **III** (2003), 2269–2272.
- [12] <http://www.speech.kth.se/multimodal/vocaltract.html>
- [13] I. S. Pandzic and R. Forchheimer (Eds.). *MPEG-4 Facial Animation. The Standard, Implementation and Applications*. Wiley, 2002.
- [14] G. Bailly. Visual speech synthesis. *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis (SSW4)*, 2001.
- [15] T. Ezzat and T. Poggio. Visual Speech Synthesis by Morphing Visemes. *International Journal of Computer Vision*, **38** (2000), 45–57.
- [16] C. Bregler, M. Covell, and M. Slaney. Video Rewrite: Driving Visual Speech with Audio. *Proceedings of the Conference on Computer Graphics and Interactive Techniques*, Los Angeles, (1997), 353–360.
- [17] S. Fagel. Video-realistic synthetic speech with a parametric visual speech synthesizer. *Proceedings of the INTERSPEECH*, Korea, 2004.
- [18] A. Hunt and A. Black. Unit selection in a concatenative speech synthesis system using a large speech database. *ICASSP-96*, vol. **1** (1996), 373–376.
- [19] U. Musti, V. Colotte, A. Toutios, and S. Ouni. Introducing visual target cost within an acoustic-visual unit-selection speech synthesizer. *Proceedings of the AVSP-2011*, 49–55.
- [20] W. Mattheyses, L. Latacz, and W. Verhelst. Auditory and photo-realistic audiovisual speech synthesis for Dutch. *Proceedings of the AVSP-2011*, 55–60.
- [21] S. Fagel. Joint audio-visual units selection the JAVUS speech synthesizer. *Proceedings of the International Conference on Speech and Computer*, St. Petersburg, 2006.
- [22] P. Wu, D. Jiang, H. Zhang, and H. Sahli. Photo-realistic visual speech synthesis based on AAM features and an articulatory DBN model with constrained asynchrony. *Proceedings of the AVSP-2011*, 61–66.
- [23] A. Eek and E. Meister. Simple perception experiments on Estonian word prosody: foot structure vs. segmental quantity. I. Lehiste and J. Ross (Eds.), *Estonian Prosody: Papers from a Symposium*, (1997), 71–99.
- [24] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. ELAN: a Professional Framework for Multimodality Research. *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*, 2006.
- [25] S. Ouni, D. W. Massaro, M. M. Cohen, K. Young, and A. Jesse. Internationalization of a Talking Head. *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS'03)* (2003).
- [26] S. Fagel. MASSY Speaks English: Adaptation and Evaluation of a Talking Head. *Proceedings of INTERSPEECH*, Brisbane, 2008.
- [27] M. Mihkla, E. Meister. Eesti keele tekst-kõne-sntees. *Keel ja Kirjandus*, (2002) **45(2)**, 88–97; **45(3)**, 173–182.
- [28] [12] The MBROLA Project, URL <http://tcts.fpms.ac.be/synthesis>.
- [29] Virtual Reality Modeling Language <http://www.vrml.org/>
- [30] M. M. Cohen and D. W. Massaro. Modeling Co-articulation in Synthetic Visual Speech. Magnenat Thalmann, N., Thalmann, D. (Eds.), *Models and Techniques in Computer Animation*, Springer-Verlag, Tokyo, (1993), 139–156.
- [31] E. Meister and L. Meister. Multimodal Corpus of Speech Production: Work in Progress. *Proceedings of the Fifth International Conference on Human Language Technologies — The Baltic Perspective*, Tartu, Estonia, October 4–5, 2012 (this volume).
- [32] M. Mihkla, L. Piits, T. Nurk, and I. Kiissel. (2008). Development of a Unit Selection TTS System for Estonian. *Proceedings of the Third Baltic Conference on Human Language Technologies*, (2008), 181–187.