Human Language Technologies – The Baltic Perspective
A. Tavast et al. (Eds.)

doi:10.3233/978-1-61499-133-5-146

Multimodal Corpus of Speech Production: Work in Progress

Einar MEISTER [1] and Lya MEISTER
Institute of Cybernetics at Tallinn University of Technology, Estonia

Abstract. The paper introduces work-in-progress on multimodal articulatory data collection involving multiple instrumental techniques such as electrolaryngography (EGG), electropalatography (EPG) and electromagnetic articulography (EMA). The data is recorded from two native Estonian speakers (one male and one female), the target amount of the corpus is approximately one hour of speech from both subjects. In the paper the instrumental systems exploited for data collection and recording set-ups are introduced, examples of multimodal data analysis are given and the possible use of the corpus is discussed.

Keywords. articulation, electrolaryngography, electropalatography, electromagnetic articulography, Estonian

Introduction

Speech as an acoustic signal is the result of complex activity in the human articulatory mechanism. The movements of different articulators are controlled by the brain, from the other side they are limited by built-in physiological constraints. In the field of phonetics, articulatory studies cover quite a large number of languages, but there are rather few attempts to exploit articulatory features in speech technology. This can be explained by the fact that study of articulation requires highly sophisticated measurement instruments and therefore there are very few articulatory databases available, e.g [1], [2], [3]. The examples of use of articulatory data in speech technology involve: (1) building realistic 3D vocal tract and tongue models [4]; (2) audiovisual speech synthesis [5], [6], [7]; (3) acoustic and audiovisual speech recognition [8], [9].

Articulation studies in Estonian date back to the 1970s including analysis of X-ray cinematography of steady vowels and dynamic electropalatography e.g. [10], [11], [12], [13]; no further studies have been carried out in the following decades. Recently facilities for the collection of multimodal corpora have been set up at the Laboratory of Phonetics and Speech Technology, IoC and the collection of the corpus of Estonian speech production has been initiated. The main aim of the corpus is to serve as a resource for diverse research tasks in phonetics and speech technology.

In this paper we give an overview of the instruments and methods of multimodal data collection and describe our work-in-progress on an articulatory-phonetic corpus of Estonian speech production.

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

1. Instrumental Techniques

Several instrumental techniques and systems have been developed and exploited for studies of articulatory processes during speech production including e.g. cinefluorography, electromyography, aerometry, endoscopy, photoglottography, tensometry, electrolaryngography, electropalatography, magnetic resonance imaging, ultrasonography, electromagnetic articulography. Our facilities for the collection of multimodal articulatory data include three systems – electroglottography, electropalatography and electromagnetic articulography.

1.1. EGG

Electroglottography (EGG) is a non-invasive technique to register laryngeal movements during speech production. Two electrodes are positioned on both sides of the thyroid cartilage and a weak voltage is passed from one electrode to the other. The change in electrical impedance across the throat depends on the contact variations between the vocal folds. These variations are recorded synchronously with the speech signal captured by a microphone. The EGG technique provides the most accurate data on the physical measures of voice fundamental frequency and larynx movements. Our EGG system is the Laryngograph Processor by Laryngograph Ltd[2], UK.

1.2. EPG

Electropalatography (EPG) is a method to study the timing and location of tongue contact with the hard palate during continuous speech. During the data capture a speaker has to wear an artificial palate; on the surface of the palate are located 62 silver contacts which register the tongue-palate contact during articulation (Figure 1). The contact data is transferred via the EPG scanner to a PC serial port interface.

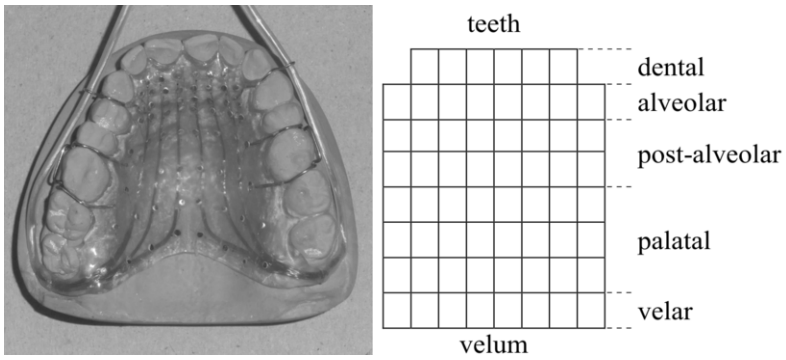

Figure 1. The EPG palate and the tongue contact areas.

The data capture process is administered with appropriate software allowing live display of the tongue-palate contact patterns and their time-synchronous recording along with the acoustic and/or laryngograph signal. Our WinEPG system is provided by Articulate Instruments Ltd[3], UK.

[2]http://www.laryngograph.com
[3]http://www.articulateinstruments.com

1.3. EMA

Electromagnetic Articulograph (EMA), The Wave Speech Research System[4] by Northern Digital Inc, Canada, is a non-line-of-sight motion capture system specifically designed for tracking speech related articulatory movements and articulatory kinematics. It enables real-time 3D data capture from 16 sensors in a measurement volume of 50x50x50 cm, along with synchronized audio. In articulatory studies the sensors are typically glued onto the tongue, the lips, the front teeth and the jaw, with one sensor glued onto the bridge of the nose acting as a reference to compensate for head movement.

Depending on the research task different systems can be simultaneously exploited in data collection; however, the synchronization of data streams from different instruments is an issue still to be solved.

2. Corpus Design and Recording Set-Ups

The multimodal corpus of speech production is aimed at studying dynamic articulatory patterns and acoustic-articulatory mapping in native Estonian speech; a further goal is to investigate the use of articulatory data in speech technology.

The text corpus compiled for data collection includes CVCV nonsense words (where V and C represent all Estonian vowels and consonants) and short sentences. Two recording set-ups have been used: (1) EGG + EPG + audio (Figure 2, left), (2) EMA + audio (Figure 2, right); in both set-ups the same text corpus has been read by two subjects (1 male, 1 female). The audio data is recorded with a close-talk microphone at a sampling frequency of 22.1 kHz.

In the EMA set-up three sensors are attached to the tongue (tongue tip, tongue blade, tongue dorsum), eight sensors to the lips (three on both upper and lower lips, two on lip corners), one to the jaw, and one sensor to the bridge of the nose.

The data collection is in progress, the target amount is one hour of speech from two subjects recorded in the two set-ups. Currently no other corpora providing data on Estonian speech production exist.

Figure 2. Left: the subject wearing electrodes of the EGG system, the EPG palate, and a close-talking microphone in the recording set-up 1; right: the position of EMA electrodes in the recording set-up 2.

[4]http://www.ndigital.com/lifesciences/products-speechresearch.php

3. Examples of Multimodal Data Analysis

The EPG and EMA systems are supplied with commercial software packages – Articulate Assistant and Wave Front/Wave View, correspondingly; for the analysis of laryngograph signals the free program EFXHIST[5] is available. For the processing and analysis of multimodal data recorded in different set-ups several software packages e.g. WaveSurfer[6], MATLAB MoCap Toolbox[7] and R[8] have been investigated. However, an optimal software environment enabling synchronization, visualization, editing, labelling, animation and analysis of data recorded from different instruments still needs to be developed.

3.1. EGG Data

Synchronously recorded speech and laryngograph signals can be analysed with the EFXHIST program (Figure 3). It finds the locations of every pitch period in the recording using the laryngograph signal and calculates statistics of voicing, fundamental frequency, closed quotient, jitter, shimmer, etc. These parameters provide measures related to the vocal fold functions and thus are widely exploited in clinical applications, see e.g. [14]. It has been found that the NAQ-feature (normalised amplitude quotient characterising the closed phase of the glottal pulse) correlates highly with the emotional state of a speaker [15] and can be used in the analysis of emotional speech.

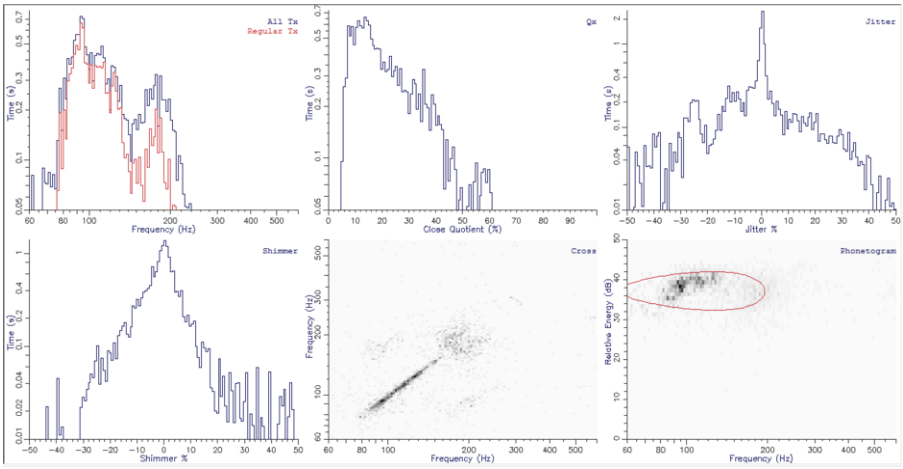

Figure 3. Different measures of voice quality calculated for one minute read speech of a male speaker. Top-left: Tx – distribution of all (blue) and regular (red) pitch periods according to their fundamental frequency; top-mid: Qx – a histogram of the closed quotient values; top-right: Jitter – a histogram of jitter (measure of period-to-period fluctuation in duration) values; bottom-left: Shimmer – a histogram of shimmer (measure of period-to-period fluctuation in speech signal amplitude) values; bottom-mid: Cross – a scatter plot of pairs of adjacent pitch periods, plotted according to the fundamental frequency of each period; bottom-right: Phonetogram – a scatter plot of the frequency and energy corresponding to each pitch period.

[5] http://www.phon.ucl.ac.uk/resource/sfs/efxhist/
[6] http://www.speech.kth.se/wavesurfer/
[7] https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mocaptoolbox/
[8] http://www.r-project.org/

3.2. EPG Data

In Estonian sounds tongue-palate contact occurs in velar and denti-alveolar plosives /k/ and /t/, alveolar and post-alveolar fricatives /s/ and /š/, alveolar nasal /n/, alveolar lateral /l/, and palatal approximant /j/ as well as in high and mid-high front vowels /i/, /ü/, /e/, /ö/. The four consonants /l, s, n, t/ have palatalised counterparts /l', s', n', t'/, correspondingly. The palatalisation is not revealed in the Estonian orthography, thus orthographically identical words can form lexical minimal pairs distinguished by palatalisation, like e.g. *palk* – /palkːk/ 'wage' vs. /palʼkːk/ 'timber'; *mats* – /matːs/ 'clout', 'whack' vs. /matʼːs/ 'boor'; *kann* – /kanːn/ 'kettle' vs. /kanʼːn/ 'flower', 'toy'. Palatalised consonants do not appear in word-initial position, they occur in the intervocalic position between the first and second syllables or at the end of words, e.g. /patːte/ 'sin', part.pl. – /patʼːte/ 'stalemate', part.pl.; /pat:t/ sin, nom.sg. – /patʼ:t/ 'stalemate', nom.sg.

EPG provides essential data for the comparing of palatalised and non-palatalised consonants revealing that the place of articulation of palatalised counterparts is characterised by a larger front and lateral contact area (Figure 4); the data recorded during a sentence provides dynamic characteristics of palatalisation (Figure 5).

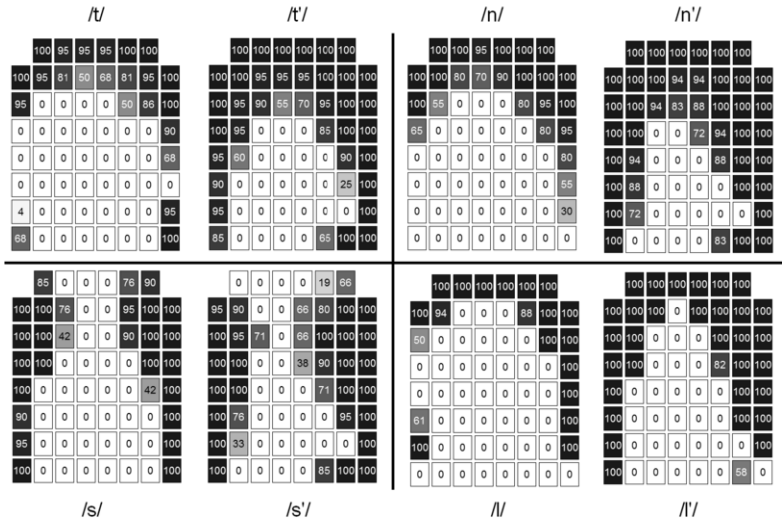

Figure 4. Palatograms of non-palatalised and palatalised Estonian consonants in intervocalic position of a CVCCV nonsense word. The numbers on the palatograms correspond to the percentage of contact time relative to the duration of a consonant for each electrode.

In addition to phonetic research, EPG has been effectively used in clinical applications as articulatory visual feedback tool for the diagnosis and treatment of several articulation disorders [16], [17]. The visualisation of EPG patterns could be useful also in foreign language pronunciation training; however, the need for individual artificial palates makes it not an easily accessible opportunity.

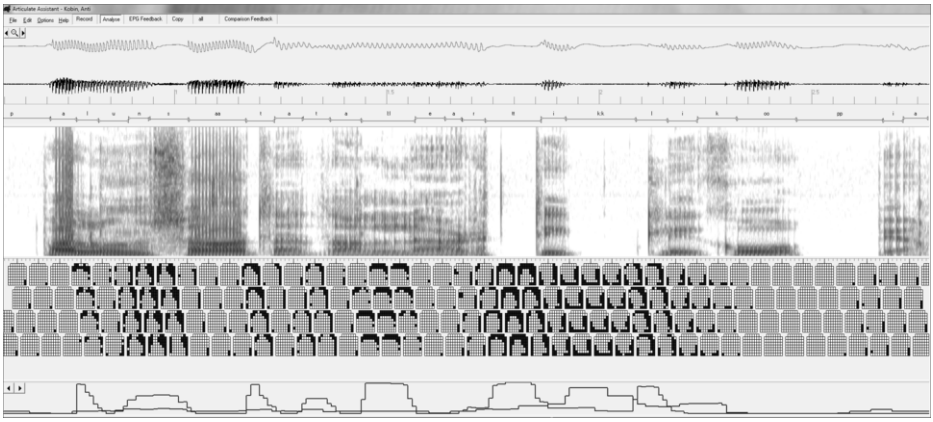

Figure 5. Multimodal data representation of the sentence "Palun saada talle artikli koopia". From top to bottom: EGG signal, speech signal, phone-level segmentation, spectrogram, palatograms, curves of denti-alveolar and velar contact.

3.3. EMA Data

EMA data recorded synchronously with audio provides information on the movements of the main articulators – tongue, lips and jaw. The data captured by the EMA system (X, Y, Z and angular coordinates of all sensors sampled at 200 Hz and synchronised with audio) is stored in a simple TSV (Tab Separated Values) format allowing easy import to Matlab or R for further processing. Figure 6 provides an illustration of EMA data showing audio signal and vertical movements of sensors attached to the upper lip, the tongue tip and the lower lip.

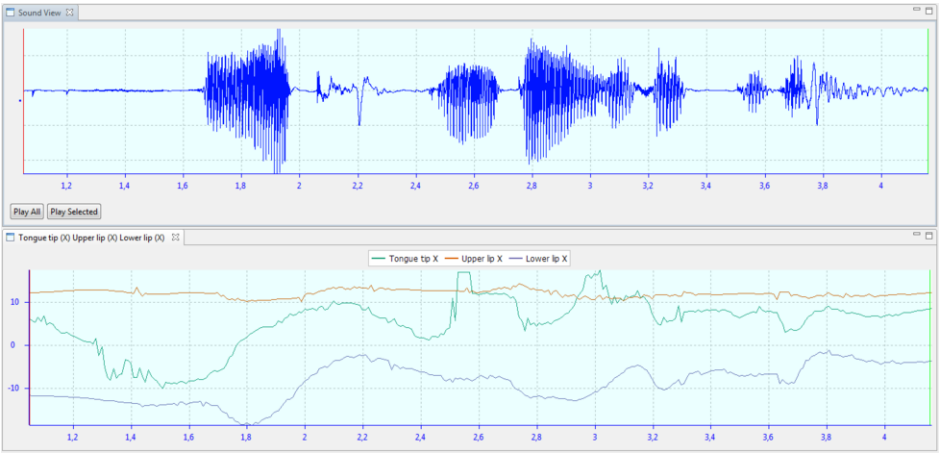

Figure 6. An illustration of EMA data in the sentence "Aeg on teele sättida". Speech signal in top window, the movement trajectories of the upper lip, the tongue tip and the lower lip in bottom window.

4. Exploiting Speech Production Data

The multimodal corpus of Estonian speech production represents a language resource to be used for multiple research tasks in phonetics and speech technology.

In phonetics, the use of the corpus will be rather traditional, i.e. it can be used for the description of articulatory patterns of Estonian speech sounds and for studies of co-articulation and acoustic-articulatory mapping. However, compared to data capture systems available in 1970s for earlier studies [10] – [13], the current corpus will provide more precise, reliable and multifaceted data.

Exploiting the corpus in speech recognition is certainly more challenging since it needs the development of new approaches and techniques for acoustic-articulatory modelling. Several authors have admitted that despite solid knowledge about speech production mechanisms accumulated in decades of research, the acoustic modelling for automatic speech recognition currently uses very little of this knowledge and no good alternative to the standard speech signal representation (i.e. MFCCs) is available (see e.g.[9]). In recent years, some progress has been made in integrating articulatory features into acoustic models and the scope of automatic speech recognition has been extended from audio-only to audiovisual speech recognition (e.g. [9], [18], [19], [20], [21]). Most of the studies have found that by combining visual and audio speech features it is possible to achieve better performance than using acoustic features only.

Using articulatory data in speech synthesis has been more successful resulting in different models of audiovisual speech synthesis (e.g. [5], [6]), especially in the case of expressive speech (e.g. [22]).

The corpus will be exploited in the development of a talking head for the Estonian audiovisual speech synthesis [23] and it will make it possible to initiate studies on Estonian multimodal speech recognition.

5. Summary

The main goal of multimodal articulatory corpus is to serve as a language resource for phonetic studies on Estonian speech production and to extend Estonian speech technology development to the synthesis and recognition of multimodal speech. The target amount of the corpus is one hour of speech recorded from two native subjects and it will be available via the Center of Estonian Language Resources.

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] A. Wrench and W. Hardcastle. A multichannel articulatory speech database and its application for automatic speech recognition. *Proceedings of the 5th Speech Production Seminar: Models and Data,* München, Germany, (2000), 305–308.

[2] M. Grimaldi, B. Gili Fivela, F. Sigona, M. Tavella, and P. Fitzpatrick. New technologies for simultaneous acquisition of speech articulatory data: 3D articulograph, ultrasound and electroglottograph. *Proceedings of LangTech 2008*, Rome, Italy, 2008.

[3] K. Richmond, P. Hoole, and S. King. Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. *Proceedings of Interspeech'2011*, Florence, Italy, (2011), 1505–1508.

[4] http://www.speech.kth.se/multimodal/vocaltract.html

[5] S. Fagel and C. Clemens. An Articulation Model for Audiovisual Speech Synthesis – Determination, Adjustment, Evaluation. *Speech Communication* **44** (2004), 141–154.

[6] 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.

[7] S. Ouni, M. M. Cohen, and D. W. Massaro. Training Baldi to be multilingual: A case study for an Arabic Badr. *Speech Communication*, **45(2)** (2005), 115–137.

[8] J. Frankel and S. King. ASR – articulatory speech recognition. *Proceedings of Eurospeech 2001*, Aalborg, Denmark, (2001), 599–602.

[9] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester. Speech production knowledge in automatic speech recognition. *Journal of the Acoustical Society of America*, **121(2)** (2007), 723–742.

[10] A. Eek, R. Haavel, O. Künnap, M. Remmel, and M. Veigel. A technique of dynamic palatography in application with a computer. *Estonian Papers in Phonetics – EPP*, Tallinn, (1973), 9–17.

[11] A. Eek. Articulation of the Estonian sonorant consonants. I-III. *Eesti NSV Teaduste Akadeemia Toimetised. Ühiskonnateadused* **19(1)** (1970), 103–121; **19(3)** (1970), 296–310; **20(2)** (1971), 173–191.

[12] A. Eek. Articulation of the Estonian sonorant consonants. IV-V. *Soviet Fenno-Ugric Studies* **7(3)** (1971), 161–168; **7(4)** (1970), 259–268.

[13] A. Eek. Observations in Estonian palatalization: an articulatory study. *Estonian Papers in Phonetics – EPP*, Tallinn, (1973), 18–37.

[14] A. Fourcin and E. Abberton. Hearing and phonetic criteria in voice measurement: Clinical applications. *Logopedics Phoniatrics Vocology* (2007), 1–14.

[15] M. Airas and P. Alku. Emootioiden vaikutus äänilähteeseen lyhyissä vokaalisegmenteiss NAQ-parametrin kuvaamana. *Fonetiikan päivät 2004*, Oulu (2004), 40–43.

[16] F. Gibbon and S. Wood. Visual feedback therapy with electropalatography (EPG) for speech sound disorders in children. In L. Williams, S. McLeod and R. McCauley (Eds.) *Interventions in Speech Sound Disorders in Children*. Brookes: Baltimore (2010), 509–536.

[17] M. J. AcAuliffe and E. C. Ward. The use of electropalatography in the assessment and treatment of acquired motor speech disorders in adults: Current knowledge and future directions. *NeuroRehabilitation* **21** (2006), 189–203.

[18] P. K. Ghosh and S. Narayanan. Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion. J. Acoust. Soc. Am. Express Letters **130 (4)**, (2011), 251–257.

[19] F. Metze. *Articulatory Features for Conversational Speech Recognition.* PhD thesis, Fakultät für Informatik der Universität Karlsruhe (TH), Karlsruhe, Germany, 2005.

[20] M. Hasegawa-Johnson, K. Livescu, P. Lal, and K. Saenko. Audiovisual speech recognition with articulator positions as hidden variables. *Proceedings of ICPhS,* (2007), 297–302.

[21] K. Livescu et al. Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU summer workshop. *Proceedings of ICASSP*, 2007.

[22] A. W. Black, H. T. Bunnell, Y. Dou, P. K. Muthukumar, F. Metze, D. Perry, T. Polzehl, K. Prahallad, S. Steidl, and C. Vaughn. Articulatory features for expressive speech synthesis. *Proceedings of ICASSP*, 2012.

[23] E. Meister, S. Fagel, R. Metsvahi. Towards Audiovisual TTS in Estonian. *Proceedings of the Fifth International Conference on Human Language Technologies — The Baltic Perspective*, Tartu, Estonia, October 4–5, 2012 (this volume).