Integration of Ohman and Rule-based Coarticulation Models for Visualization of Pure Lithuanian Diphtongs

I. Mazonaviciute[1], R. Bausys[1], A. Kriukovas[1]
[1]*Department Department of Graphical Systems, Vilnius Gediminas Technical University, Saulėtekio al. 11. SRKII - 608, 10223 Vilnius, Lithuania, phone: + 370 527 448 48*
ingrida.mazonaviciute@vgtu.lt

***Abstract*—Visualization methodology for pure Lithuanian diphthongs that are influenced by neighbouring phonemes are presented. Ohman and Rule-based coarticulation control models are integrated. Rules for Lithuanian diphthong animation are defined based on Lithuanian phonology. Most of them can be realized using Ohman control model, others use linear interpolation of expressiveness parameters. The proposed coarticulation model can be successfully applied to define expressiveness coefficients in viseme-driven speech animation engines.**

***Index Terms*—Speech analysis, coarticulation, phoneme, viseme, speech animation.**

I. INTRODUCTION

Movements of vocal organs which are involved in the production of speech sounds (articulators) significantly improve the understanding of acoustic signal. So, computer generated 3D human head models, specified to animate synthesized or natural speech (talking heads) recently became an important part of human-computer communication. Talking heads can be employed for e-consulting services: virtual secretary, WEB navigator or virtual agent who is responsible for information conveying to user in a Smart Ecological and Social Apartments (SESA). Also talking heads are widely used in e-learning technologies for the correct presentation of the sound pronunciation [1] or applied in movie, advertising and computer game industries.

Invention of new speech animation engine is time consuming task and also requires considerable financial support and specific knowledge. Translingual speech animation can be applied to save resources and to integrate a talking head based on phonetics in foreign language. For instance, visual similarity of Lithuanian and English phonemes was used to animate Lithuanian speech using English speech animation engine [2].

However, every language is unique and has specific phonetic rules for speech production. Moreover, a coarticulation effect that refers to the situation in which a conceptually isolated speech sound is influenced by neighbouring segments has a high influence on visual speech. Coarticulation control models [3] should be applied to govern the articulatory movements for a given phonetic target specification. Typically the coarticulation control models can be expressed by a sequence of time labelled phonemes, including stress and phrasing markers.

Manuscript received March 11, 2012; accepted May 14, 2012.

II. COMPARISON OF COARTICULATION CONTROL MODELS

Different researchers identified various coarticulation control models including Rule-based [4], Cohen-Massaro [5], Ohman [6] and Artificial Neural Network (ANN) [7] models. Rule-based (parametric) coarticulation models use a set of explicit rules to model steady-state properties of pronounced phonemes and parametrically control how these phonemes are fused into connected speech. For instance, Beskow [4] proposed rule–based model, where each phoneme is assigned to a target vector of articulatory control parameters. Some parameter values can be left undefined to allow these targets to be influenced by coarticulation. If a target is left undefined, the value is inferred from context using interpolation. For example, the lip rounding parameter in V_1CCCV_2 utterance (vowel V_1 is unrounded, V_2 – rounded) is unspecified for the consonants C, so consonant targets are determined from the vowel context by linear interpolation from V_1, to V_2.

Most rule-based animation engines are structured on formal linguistic theory, so implementation of rule-based model is limited by people‘s incomplete understanding of coarticulation effect and their inability to build a full set of rules for phonemes, that are influenced by neighbours.

Data-driven animation engines commonly exploit Cohen-Massaro, Ohman [6] and ANN coarticulation control models. Cohen-Massaro and Ohman models are associated with speech production theory. Artificial neural networks (ANN) [7] must be trained to predict articulatory parameter values on a frame-by-frame basis.

Cohen-Massaro [5] uses dominance function that gradually increases up to a peak value and then decreases.

This function is employed to model articulatory gestures. However, certain targets, such as the closure in a bilabial stop, cannot be achieved with this model.

Ohman coarticulation control model modified by Reveret [8] defines coarticulation between two vowels. The vowel

track $v(t)$ is formed by interpolation between fully expressed vowel targets (visemes). So, this model can be applicable to track parameters of the visemes that appear in phonetic structures V_1CV_2, V_1CCV_2, V_1CCCV_2 etc. A consonant is specified by a target value c, a coarticulation factor w_c and a function $k(t)$ that dictates the temporal blending of vowel track and consonant target. Both w_c and $k(t)$ are in the interval [0, 1]. The trajectory of a given articulatory parameter over arbitrary phoneme sequences can be described as

$$z(t) = v(t) + \sum_{i \in C} w_{ci} k_i(t)(c_i - v(t)), \quad (1)$$

where C is the set of all consonants in the analyzed phonetic structure.

The vowel track $v(t)$ is formed by temporally blending successive fixed vowel targets a_i, according to the function

$$v(t) = \frac{\sum_{j=1}^{N} a_j b_j(t)}{\sum_{j=1}^{N} b_j(t)}, \quad (2)$$

where N is the number of vowels in the analysed utterance and $b_j(t)$ is the blend function of the j^{th} vowel in the utterance. Cubic function $b_j(t)$ has the value $b_j(t) = 1$ at the centre of vowel j and the value $b_j(t) = 0$ at the centre of the preceding $(j-1)$ and following $(j+1)$ vowel.

Since there is no intervocalic coarticulation in the Ohman model, the blending function $b_j(t)$ can be also applied to consonants that are between vowels.

Perceptual intelligibility experiment compared coarticulation control models together with an audio-alone condition (Table I) [3].

TABLE I. SUMMARY OF INTELLIGIBILITY TEST OF VISUAL SPEECH SYNTHESIS CONTROL MODELS [3].

Control model	% keywords correct
Audio only	62.7
Cohen-Massaro	74.8
Ohman	75.3
ANN	7.,8
Rule-based	81.1

The results confirm that all coarticulation control models give significantly increased speech intelligibility over the audio-alone case (Table 1). It also demonstrates that rule-based model has the highest intelligibility score and Ohman is more effective than Cohen-Massaro. So, in our research we propose to integrate Rule-based and Ohman coarticulation models to visualize pure Lithuanian diphthongs.

III. COARTICULATION ANALYSIS OF LITHUANIAN DIPHTHONGS

The term diphthong refers to two adjacent vowel sounds occurring within the same syllable. Technically, a diphthong is a vowel with two different targets. This means that tongue and lips move during the pronunciation of this vowel.

Languages differ in the length of diphthongs. Diphthongs typically behave like long vowels in languages with phonemically short and long vowels. Lithuanian language is a good example of this case [9]. Besides, languages differ in the count of diphthongs (10 in British English [10], 6 in Dutch [11], etc.).

There are two types of Lithuanian diphthongs: 9 pure (*Vowel-Vowel structure (VV)*) and 20 mixed diphthongs, that are made of vowels “a”, “e”, “i”, “u” and consonants “l”, “m”, “n”, “r” and has the *Vowel-Consonant structure (VC)* [12]. Only pure Lithuanian diphthongs (*ai, au, ei, ui, ie, uo, eu, oi, ou*) will be analysed in our research. They regularly appear in Lithuanian words (e.g. *miegas, saulė, eisena*), so visualization of these diphthongs requires additional attention; especially when English speech animation engine is used to animate Lithuanian speech. Architecture of Lithuanian speech animation engine was presented in [2].

Stressing of the diphthong highly influence visualization of the speech since stressed syllable is more expressive then others. Besides, position of the accented phoneme strongly influences appearance of neighbouring phonetic segments. English diphthongs are always stressed with the falling accent and Lithuanian diphthongs can be stressed with rising accent, too. Thus, three situations of Lithuanian diphthong stressing can be distinguished:

1) Diphthong is stressed with falling accent (ái). Falling diphthong starts with a vowel quality of higher importance (higher pitch or volume) and ends in a semivowel with less prominence. Examples of falling diphthongs: láimė, áugti, léisti;

2) Diphthong is stressed with rising accent (aĩ). Rising diphthong begins with a less prominent semivowel and end with a more prominent full vowel. Examples: eĩti, šiẽnas;

3) Diphthong is in the unstressed syllable (e.g. traukinys, gžuolas).

People prepare themselves for pronunciation of the next phoneme during articulation of the current phoneme. So, the second group of features that influence visualization and expressiveness of Lithuanian diphthongs includes its location in the word and its neighbours.

Three cases of diphthong location in the word can be distinguished:

*1) End of word is one letter to the right of the current diphtong (VVC structure) (e.g. “tak**ai**s);*

*2) End of word is in the diphthong (CVV structure) (e.g. „tak**ai**“);*

*3) End of word is somewhere else (VVCV, VVCCV etc. structures) (e.g. „t**ai**ka, **ai**tvaras“).*

Diphthong position in the word together with information about its stressing are analysed to investigate their influence for speech animation. Rules for Lithuanian diphthong visualization are included in the framework for Lithuanian diphthong visualization (Fig. 1).

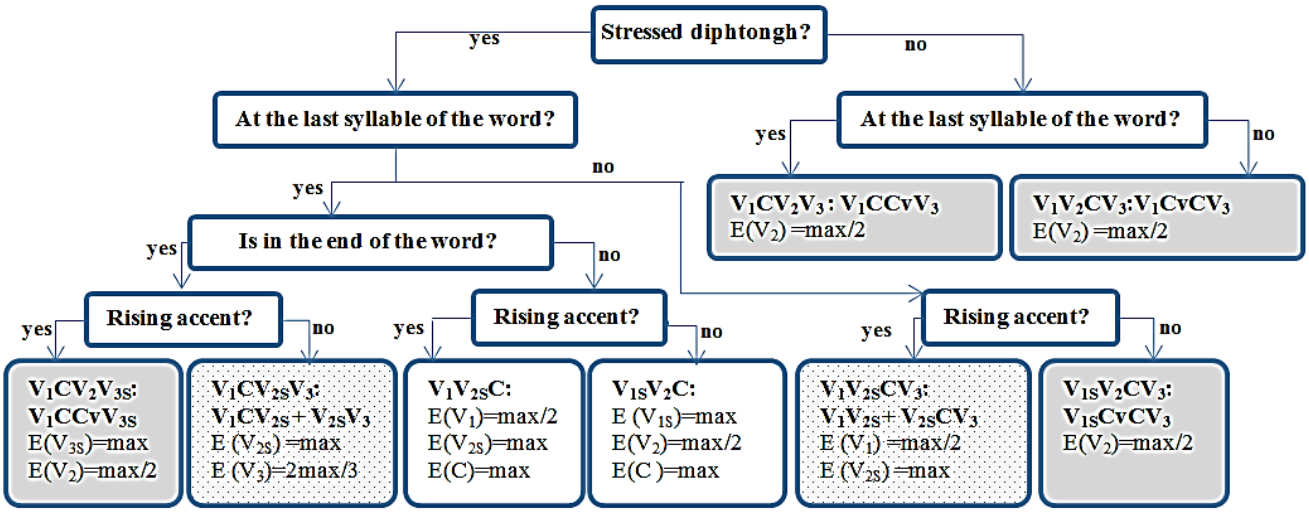

Fig. 1. Framework for pure Lithuanian diphthong animation. Dotted and greyed rules define situation when Ohman coarticulation control model can be applied and white rules define state, where it cannot be done.

IV. PROPOSED FRAMEWORK FOR PURE LITHUANIAN DIPHTHONG VISUALIZATION

Rules for pure Lithuanian diphthong visualization (Fig. 1) can be divided into two main groups: rules where adapted Ohman coarticulation control model can be applied (dotted and greyed rules) and those, where it cannot be done (white rules).

First group of rules uses Vowel-Vowel-Consonant-Vowel (VVCV) or Vowel-Consonant-Vowel-Vowel (VCVV) utterances for diphthong visualization, where structure VV defines a pure Lithuanian diphthong.

Ohman coarticulation control model proposed earlier does not define coarticulation between two consonants. So, the influence of analysed consonant must extend no further than to the peak of the preceding or following gesture. Moreover, Ohman model [8] is designed for VCV, VCCV or VCCCV phonetic utterances, therefore its application for VVCV and VCVV utterances must be considered separately. So, in this paper we propose technique, how VVCV and VCVV utterances can be visualized using Ohman coarticulation model designed for VCCV phonetic structure.

Lithuanian diphthong can be stressed in three ways (falling accent, rising accent or non-stressed), so the influence of the stressed vowel V_S for the appearance of non-accented vowel of diphthong should be analysed. It was stated earlier, that falling diphthong ($V_{1S}V_2$) starts with a vowel quality of higher importance and ends in a semivowel with less prominence. It means that semivowel ($V_{semi} = V_2$) is visually much less expressed too. Therefore we define that supreme expressiveness coefficient $E(V_{semi})$ for viseme of semivowel is equal to the half of vowel's V_2 maximum expressiveness

$$E(V_{semi}) = \frac{E(V_{2\max})}{2}. \quad (3)$$

The rising diphthong V_1V_{2S} begins with a semivowel and ends with a more prominent full vowel, so $V_{semi} = V_1$. In the meantime, we treat the non-stressed (V_1V_2) diphthong as duet of two semivowels ($V_1V_2 = V_{1semi}V_{2semi}$).

Vowel articulation strongly influences pronunciation of neighbouring phonemes [10]. Consonants' visual expressiveness is much lower, so they are highly dependable from neighbouring vowels. In the meantime semivowel has similar characteristics as consonant: it is highly influenced by stressed vowel and its expressivity is much lower. So in VVCV and VCVV utterances, we propose to treat semivowel as virtual consonant (C_V), which has the maximum expressiveness values equal to the maximum expressiveness of semivowel. This transcription gives us possibility to use Ohman coarticulation rule model for diphthongs in VCVV and VCVV syllables. For instance: syllable $V_{1S}V_2CV_3$ can be transformed into utterance $V_1C_VCV_3$, which is suitable for diphthong visualization with Ohman model.

Different adaptation of Ohman model should be done for coarticulation rules that are the dotted in Fig. 1. Analysis of the phonetic structure $V_1V_{2S}CV_3$ shows, that it should be transformed to $C_VV_{2S}CV_3$ utterance, but the whole phonetic structure is not suitable for Ohman model. On the other hand, structure $V_1V_{2S}CV_3$ can be split into two parts: V_1V_{2S} and $V_{2S}CV_3$ utterances. Ohman coarticulation rule model can be applied for $V_{2S}CV_3$ utterance and linear interpolation can be applied for V_1V_{2S} structure, where V_1 is a semi vowel. In the case, when unstressed vowel of the diphthong is the last phoneme of the word and its visual appearance is very important for speech understanding, expressiveness coefficient for viseme of this semivowel is equal to 2/3 of vowels maximum expressiveness.

Finally two white rules in Fig. 1. describe the situation, when stressed diphthong is at the last syllable of word, which ends in consonant. Ohman is not suitable, so we propose to apply linear interpolation between expressivity coefficients of these phonemes.

To estimate quality of our proposed Ohman and Rule-based control model, we've compared visemes expressivity

before and after application of our model. Results of this experiment are shown in Fig. 2. They confirm that expressiveness parameters of visemes defined by proposed coarticulation model are much more reliable to coarticulation of Lithuanian word *"**juo**das"*, which includes pure Lithuanian diphthong.

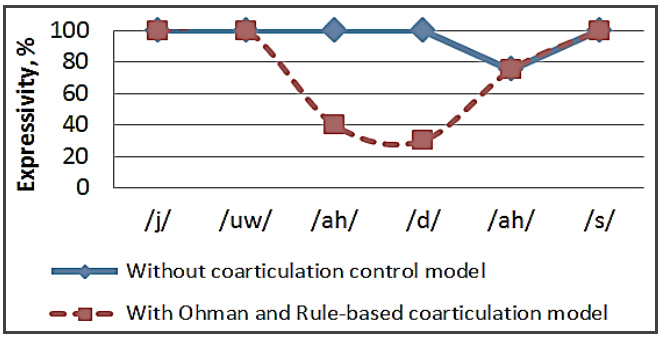

Fig. 2. Expressiveness parameter before and after application of our proposed model.

V. CONCLUSIONS

Since all humans are experts in lip reading and detects even the slightest errors during speech animation, expressive speech with integrated coarticulation rules is crucial part of any speech animation system. Ever since pure Lithuanian diphthongs regularly appear in Lithuanian words, their accurate visualization considerably improves the perspicuity of animated speech.

Eight rules for pure Lithuanian diphthong animation were defined in this paper. Six of them employ adapted Ohman coarticulation control model to define expressiveness parameter of visemes and the rest of them exploit linear interpolation between expressiveness parameter of visemes. The proposed integration of Ohman and Rule-based coarticulation models was applied in rule-based Lithuanian speech animation engine [2]. Comparison of visemes expressivity before and after integration of the proposed rules proved that intelligibility of animated Lithuanian words with pure Lithuanian diphthongs noticeably increased.

REFERENCES

[1] S. G. Pentiuc, O. A. Schipor, M. Danubianu, M. D. Schipor, I. Tobolcea, "Speech Therapy Programs for a Computer Aided Therapy System", *Elektronika ir Elektrotechnika (Electronics and Electrical Engineering*), no. 7, pp. 87–90, 2010.

[2] I. Mazonaviciute, R. Bausys, "Translingual visemes mapping for Lithuanian speech animation", Elektronika ir Elektrotechnika (Electronics and Electrical Engineering), no. 5, pp. 95–98, 2011.

[3] J. Bescow, "Trainable articulatory control models for visual speech synthesis", *Journal of Speech Technology,* vol. 4, no. 7, pp. 335–349, 2004. [Online]. Available: http://dx.doi.org/10.1023/B:IJST.0000037076.86366.8d

[4] J. Bescow, "Rule-based Visual Speech Synthesis", in *Proc. of the 4th European Conference on peech Communication and Technology (Eurospeech 95),* Madrid, Spain, 1995, pp. 299–302.

[5] M. M. Cohen, D. W. Massaro, "Modelling Coarticulation in Synthetic Visual Speech", *Models and Techniques in Computer Animation,* Springer Verlag, Tokyo, 1993, pp.139–156.

[6] S. Ohman, "Numerical model of coarticulation", *Journal of the Acoustical Society of America,* no. 41, pp. 310–320, 1967. [Online]. Available: http://dx.doi.org/10.1121/1.1910340

[7] N. Strom, "Phoneme probability estimation with Dynamic Sparsely Connected Artificial Neural Networks", *The Free Speech Journal,* vol. 1, no. 5, 1997.

[8] L. Reveret, G. Bailly, P. Badin, "Mother: a New Generation of Talking Heads Providing a Flexible Articulatory Control for Video-Realistic Speech Animation", in *Proc. of the 6th International Conference on Spoken Language Processing (ICSLP 2000),* Bejing, 2000, pp. 755–758.

[9] A. Girdenis, *Teoriniai fonologijos pagrindai.* Vilnius, 1995.

[10] P. Roach, "British English: Received Pronunciation", *Journal of the International Phonetic Association,* vol. 34, no. 2, pp. 239–245, 2004. [Online]. Available: http://dx.doi.org/10.1017/S0025100304001768

[11] Jo Verhoeven, "Belgian Standard Dutch", *Journal of the International Phonetic Association,* vol. 35, no. 2, pp. 243–247, 2005. [Online]. Available: http://dx.doi.org/10.1017/S0025100305002173

[12] P. Kasparaitis, "Lithuanian Speech Recognition Using the English Recognizer" *Informatica*, vol. 19, no. 4, pp. 505–516, 2008.