Е-списание "Образование и развитие" ISSN 2603-3577
Брой 7, 02.2021 г.

СИНХРОНИЗАЦИЯ НА УСТНИТЕ С ВЪЗПРОИЗВЕЖДАНАТА РЕЧ (LIPSYNCH) ПРИ ИГРОВИ ПЕРСОНАЖИ В ДИГИТАЛНИ ИГРИ, НАСОЧЕНИ КЪМ ОБУЧЕНИЕТО

Ивайло Ив. Буров

Резюме: *В статията е направен преглед на съществуващи решения за синхронизация на устните на 3D персонажи с аудио запис на реч. Анализирани са преимуществата и недостатъците на различните решения. Публикувани са резултатите от наблюденията със собствено решение, базирано на разпознаване на форманти в реално време и са засегнати забелязаните недостатъци при този подход. Направено е кратко изложение на теоретичната част, отнасяща се до формантите.*

Ключови думи: *игрови персонаж, синхронизация на устните, 3D, форманти, висеми*

LIP SYNCHRONIZATION OF GAME CHARACTERS IN DIGITAL EDUCATIONAL GAMES (LIPSYNCH)

Ivaylo Iv. Burov

Abstract: *The paper reviews existing solutions for lip-syncing 3D characters with audio speech recordings. The advantages and disadvantages of the different solutions are analyzed. The results of observations with a proprietary solution based on real-time formant recognition are published and the noted shortcomings of this approach are addressed. A brief exposition of the theoretical part relating to formants is made.*

Keywords: *game character, lip sync, 3D, formants, visems*

Една от най-необходимите функционалности при създаването на интерактивно 3D съдържание за целите на обучението е наличието на диалози между персонажите. Това изисква създаването на система за управление на диалозите и възможност за синхронизация на възпроизвеждания диалог с анимацията на лицето (устните) на персонажа (Lip Synch).

Синхронизацията на устните с възпроизвежданата реч при диалозите е един трудоемък процес, понякога даже избягван в комерсиалните игри като Heroes Of Might And Magic 3 и др. В някои от компютърните игри в диалозите просто се прикрепват субтитри без звук и лицева анимация, в други само субтитри със звук. В игрите с високо качество, известни като AAA games (трийпъл ей игри) това се реализира от тимове,

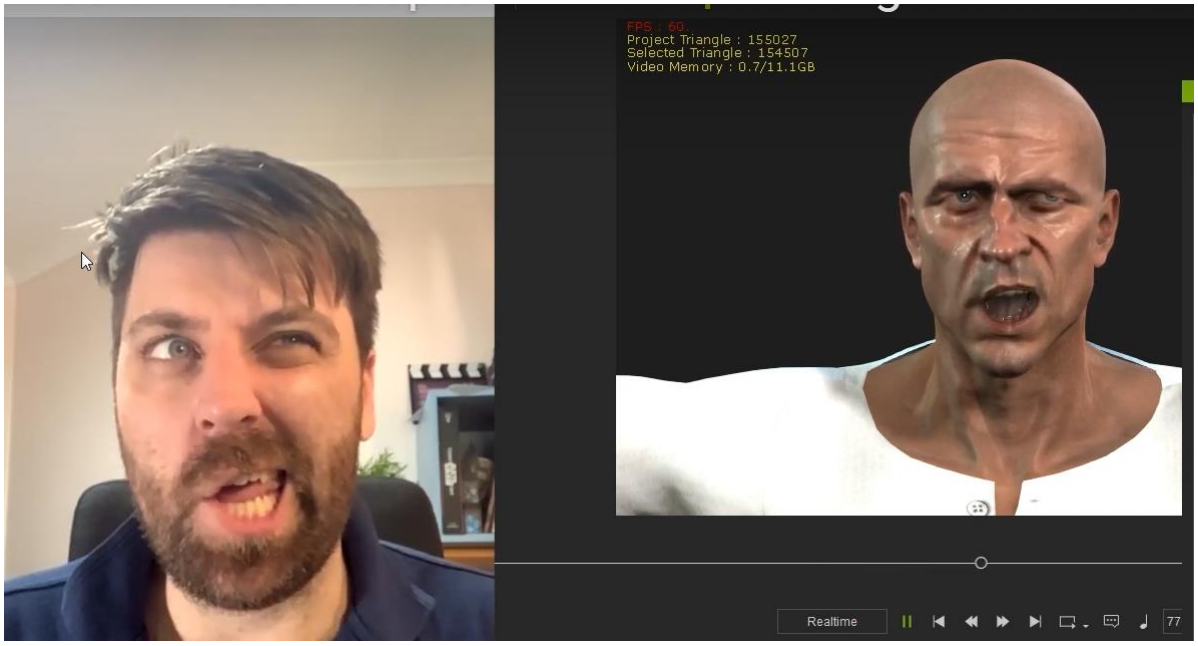
специализирани в този вид дейност. Обемът на работа се увеличава драстично, когато се изисква и многоезикова поддръжка.

Отчитайки тези фактори, при създаването на некомерсиална игра, насочена към обучението, идеята за Lip Synch поддръжка първоначално изглежда безперспективна. Съществуват обаче възможности, които са обект на описания тук експеримент.

За преценка на възможностите за осигуряване на такава функционалност, преди всичко е необходимо да бъдат разгледани известните широко разпространени методи.

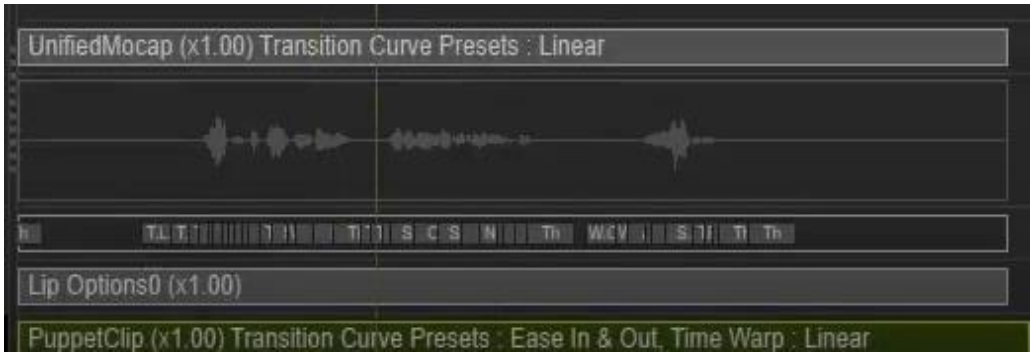
Посредством комерсиални продукти за улавяне на лицева анимация

Обикновено Lips Synch функционалност се поддържа от някои професионални комерсиални продукти за 3D анимация на персонажи или се предлага като комерсиално разширение към подобен продукт. Пример за такова разширение е Facial Motion Capture Animation (Улавяне на лицева анимация), към комерсиалния продукт iClone, където се захващат лицевите анимации на човека пред камерата, като в реално време, но с известно закъснение се прилагат към избран 3D модел (фигура 1).



Фигура 1. Улавяне на лицева анимация в комерсиалния продукт iClone

Освен улавянето на лицевите анимации се генерира и съответстваща фонограма (Фигура 2).



Фигура 2

От фонограмата се вижда, че е необходимо допълнително ръчно синхронизиране на генерираните фонеми с отделните фонограмни области, което също е времеемък процес. Още повече, всяка анимация от диалозите трябва да се записва отделно, което е по-подходящо при създаването на филмово съдържание или видеоклип, отколкото за експорт в игрови двигател. Разширението е предназначено за работа с английски език, което го прави неприложимо при диалози на български.

При създаването на игри с многомилионен бюджет за постигане на висок реализъм се използва такава технология, но в контекста на разработването на некомерсиален и свободен за обучението софтуер, този подход е нецелесъобразен.

Посредством Rhubarb Lip Sync – решение с отворен код

Друго налично решение е с помощта на инструмента rhubarb-lip-sync, като за него има и примери за ползване в игровия двигател Godot. Този инструмент се изпълнява под конзола във Windows и за определен звуков файл създава json или XML файл, съдържащ разпознати висеми със съответните им тайминги, съответстващи с тези на фонемите от звуковия файл.

Rhubarb Lip Sync се използва предимно в 2D анимацията, но се прилага и в някои 3D приложения. Използват се между шест и девет различни позиции на устата. Първите шест форми на уста (Ⓐ-Ⓕ) са основните форми на устата и абсолютният минимум, който се препоръчва. Тези шест форми на устата са създадени в студиото Hanna-Barbera за шоута като Scooby-Doo и The Flintstones. Оттогава те са се превърнали в нещо като стандарт за 2D анимация и са били широко използвани от студия като Disney и Warner Bros. В допълнение към шестте основни форми на устата, има три удължени форми: Ⓖ, Ⓗ и Ⓧ.

Ⓐ		Затворена уста за звуците "P", "B" и "M". Това е почти идентично с формата Ⓧ, но има толкова лек натиск между устните.
Ⓑ		Леко отворена уста със стиснати зъби. Тази форма на устата се използва за повечето съгласни („K“, „S“, „T“ и т.н.). Използва се и за някои гласни, като например звука „EE“ в bee.
Ⓒ		Отворена уста. Тази форма на устата се използва за гласни като "EH", както при man (човек), и "AE", както при bat (прилеп) . Използва се и за някои съгласни, в зависимост от контекста. Тази форма се използва и като междинна, при анимиране от Ⓐ или Ⓑ към Ⓓ.
Ⓓ		Широко отворена уста. Тази форма на устата се използва за гласни като „AA“, както във father (баща).

Ⓔ		Леко заоблена уста. Тази форма на устата се използва за гласни като „AO“ като off и „ER“ като bird. Тази форма се използва и като междинна при анимиране от Ⓒ или Ⓓ към Ⓕ.
Ⓕ		Набръчкани устни. Тази форма на устата се използва за "UW", както при y**ou** (вас), "OW" (УОУ), както в sh**ow** (шоу), и "W", както във **w**ay (начин).
Ⓖ		Горните зъби докосват долната устна за „F“ като във **f**or (за) и във „V“ **v**ery (много). Тази удължена форма на устата не е задължителна.
Ⓗ		Тази форма се използва за дълги звуци "L", с повдигнат език зад горните зъби. Устата трябва да е поне широко отворена, както при Ⓒ, но не толкова, колкото при Ⓓ. Тази удължена форма на устата не е задължителна.
Ⓧ		Бездействие. Използва се при паузи в речта. Съответства на позицията на устата, когато персонажът се разхожда, без да говори. Почти идентична е с Ⓐ, но с по-малко натиск между устните: За Ⓧ устните трябва да са затворени, но отпуснати. Тази удължена форма на устата не е задължителна.

Генерираният json или XML файл може да бъде въведен в игровия двигател и на база висемите и таймингите, да се приложат съответните висемни форми на персонажния модел, съответстващи на висемите от json файла. Наличните примери са с разпознаване на английски език. Пример съдържащ данни от такъв файл е представен по-долу:

```
  "metadata": {
    "soundFile": "C:\\Users\\Daniel\\Desktop\\av\\hi\\hi.wav",
    "duration": 0.47
  },
  "mouthCues": [
    { "start": 0.00, "end": 0.05, "value": "X" },
    { "start": 0.05, "end": 0.27, "value": "D" },
    { "start": 0.27, "end": 0.31, "value": "C" },
    { "start": 0.31, "end": 0.43, "value": "B" },
    { "start": 0.43, "end": 0.47, "value": "X" }
  ]
}
```

Налична е и опция за генериране на данните в XML формат. Съществуват две опции за разпознаване на речта аудио записа: pocketSphinx (за записи на английски) и фонетична (phonetic) –за записи, които не са на английски.

Първата стъпка в обработката на аудио файла се състои в определянето на какво се казва, т.е. какъв звук се произнася и в кой момент от времето.

При първата опция за английски език се използва библиотеката с отворен код за разпознаване на реч PocketSphinx, която като цяло дава добри резултати. Това е опцията по подразбиране. Недостатъкът е, че PocketSphinx разпознава само английски диалог.

Rhubarb Lip Sync се предлага и с фонетично разпознаване. При тази опция разпознаването няма да се опитва да разбере цели (английски) думи и фрази. Вместо това се разпознават отделни звуци и срички. Резултатите обикновено са по-малко прецизни

от тези от разпознавателя PocketSphinx. Предимството е, че този вариант може да се използва с езици, различни от английския.

При вариантите с използването на Rhubarb Lip Sync трябва да се акцентира на това, че разпознаването не става в реално време. Всеки звуков файл трябва да се подаде за обработка и да се използва по-късно в приложението или играта. От това следва, че ако се промени част от диалога, то висемните файлове трябва да се генерират отново. Обикновено за всяка реплика на персонаж се използва отделен файл, което означава обработка на голямо количество файлове. Когато е необходимо да се направят корекции в диалозите, което често е необходимо, процесът трябва да се повтори за всяка корекция. При мултиезикова поддръжка времето, необходимо за тези действия се умножава. Понеже резултатите при езици различни от английския не са толкова прецизни може да се потърси друг вариант като алтернатива.

Експериментално решение, посредством формантно разпознаване в реално време

В настоящия експеримент е избран друг подход – разпознаване на фонемите посредством отчитане на налични форманти от първи и втори порядък в реално време и последващо прилагане на разпознатите фонеми към определени лицеви форми (висеми). За да бъдат разбрани приложените методи е необходимо да се разгледа част от свързаната с тях теория.

В науката за речта и фонетиката, формантът е широкият спектрален максимум, който е резултат от акустичен резонанс на човешкия гласов тракт (Titze 1994)(Titze et al 2015).

В акустиката формантът обикновено се определя като широк пик или локален максимум в спектъра (Jeans 1938)(Standards Secretariat, Acoustical Society of America 1994). За хармоничните звуци според тази дефиниция, честотата на форманта понякога се приема като тази на хармоника, която е най-усилена от резонанс. Разликата между тези две дефиниции се състои в това дали формантите характеризират механизмите за възпроизвеждане на звук или самият произведен звук. На практика честотата на спектрален пик се различава леко от свързаната резонансна честота, освен когато хармониците са подравнени с резонансната честота.

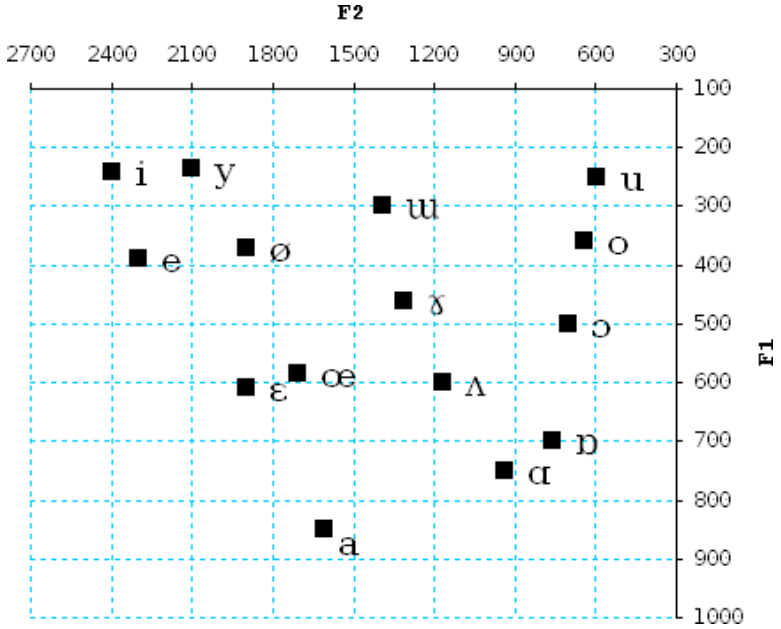
Във всяка една стая например при звук се възпроизвеждат форманти, характерни за тази конкретна стая, поради нейните резонанси, т.е. на начина, по който звукът се отразява от нейните стени и предмети. Формантите във стаята по същество се оформят посредством усилване или поглъщане на специфични честоти. По същия начин се оформят и формантите при речта, но спрямо особеностите на конкретния човешки гласов тракт.

Терминът „формант“ е въведен от немският учен Лудимар Херман през 1894 г. Според него гласната е специален акустичен феномен, зависещ от периодичното възпроизвеждане на специална частична или „формантна“, характеристика. Честотата на

форманта може да варира малко, без да променя характера на гласната. За „дълго е“ (ее или iy) например, най-нискочестотният „формант“ може да варира от 350 до 440 Hz дори при един и същ човек (McKendrick 1903).

Формантите са отличителни честотни компоненти на акустичния сигнал, произведен от реч или пеене. Информацията, от която хората се нуждаят, за да разграничат звуците на речта, може да бъде представена чисто количествено чрез определяне на пикове в честотния спектър.

Формантът с най-ниска честота се нарича F1, вторият F2, а третият F3. Основната честота или височина на гласа понякога се нарича F0, но това не е формант. Най-често първите два форманта, F1 и F2, са достатъчни за идентифициране на гласната. Връзката между възприеманото качество на гласната и първите две формантни честоти може да бъде оценена чрез прослушване на изкуствено генерирани гласни, филтрирани през двойка лентови филтри за симулиране на формант от гласов тракт). Носните съгласни обикновено имат допълнителен формант около 2500 Hz. Фигура 3 съдържа диаграма, на средните форманти на гласни букви. В таблица 1 са представени честотните характеристики на формантите от първи и втори порядък, съответстващи на гласните по международна фонетична азбука (IPA International Phonetic Alphabet). В дясната колона е стойността на честотната разлика между двата форманта.



Фигура 3. Диаграма на средните форманти на гласни

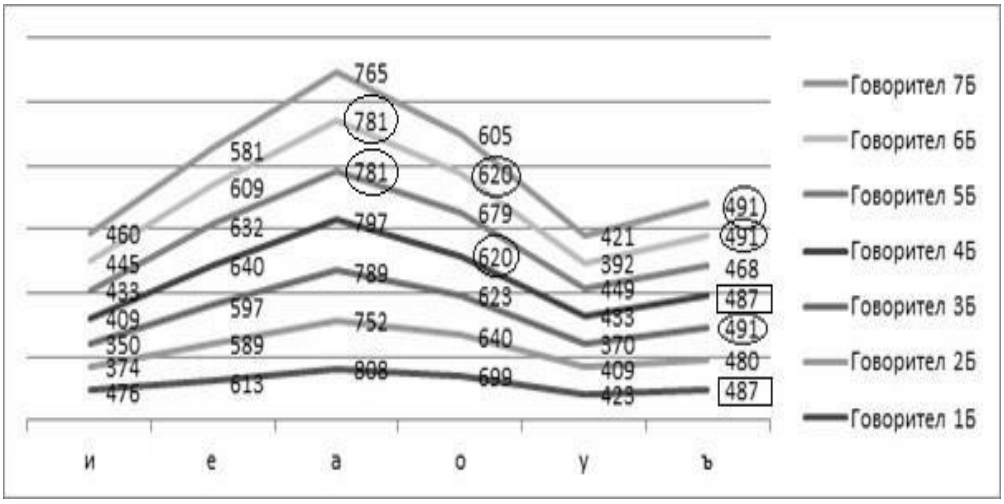
Средни форманти на гласни за мъжки глас Catford (1988)			
Гласна (IPA)	**Формант F_1 (Hz)**	**Формант F_2 (Hz)**	**Честотна разлика $F_2 - F_1$ (Hz)**
i	240	2400	2160
y	235	2100	1865
e	390	2300	1910

ø	370	1900	1530
ɛ	610	1900	1290
œ	585	1710	1125
a	850	1610	760
ɶ	820	1530	710
ɑ	750	940	190
ɒ	700	760	60
ʌ	600	1170	570
ɔ	500	700	200
ɤ	460	1310	850
o	360	640	280
ɯ	300	1390	1090
u	250	595	345

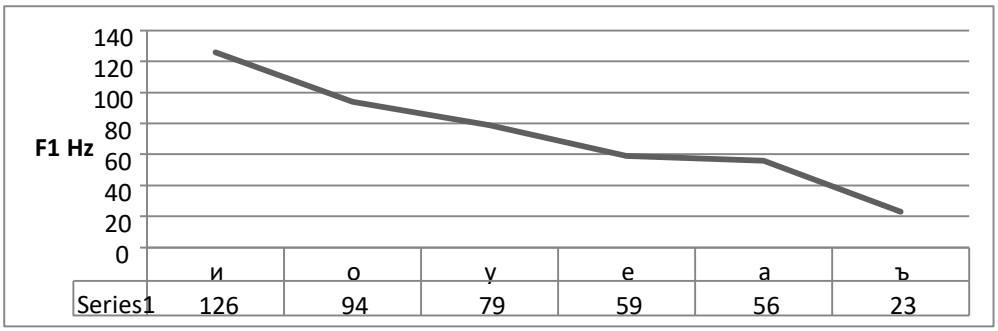
Таблица 1

Следва да се отбележи, че практически формантите не са фиксирани както посочените в таблицата, а честотните им характеристики варират в зависимост от пола или индивидуалните характеристики на говорителя.

Н. Чочева провежда изследване, като получава данни за вариативността на гласните в българския език, представени на фигура 4 и фигура 5.

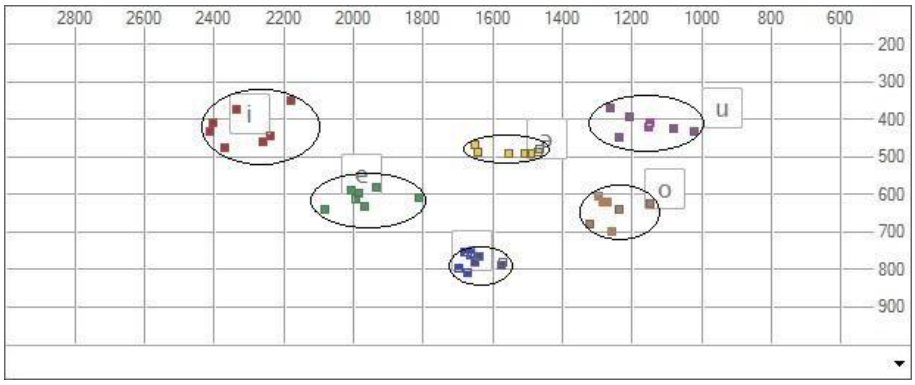


Фигура 4. Стойности на F1 на българските гласни (Чочева Н. 2014)



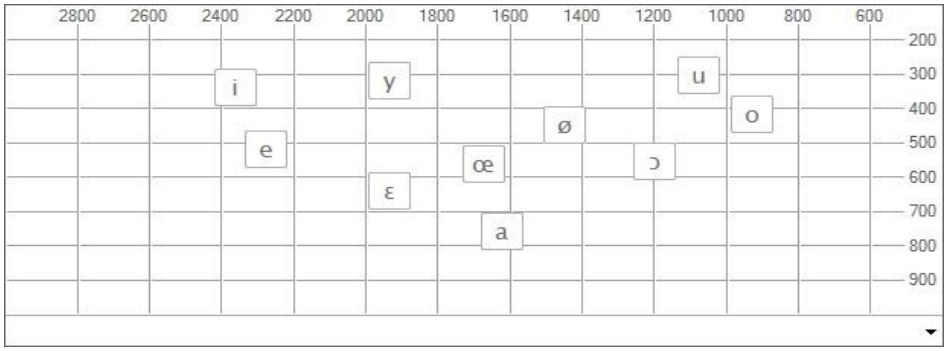
Фигура 5 Вариативност на българските гласни по стойности на F1 (Чочева Н. 2014)

Такава вариативност е отчетена и за форманта от втори порядък. Следователно за получаване на по-точен резултат е необходимо да се филтрират два честотни диапазона за определяне на всеки формант, което повишава вероятността за грешно определяне на формантите. Според същото изследване е представена и диаграма на дисперсните зони, базирана на резултатите от седем журналистки (Фигура 6).

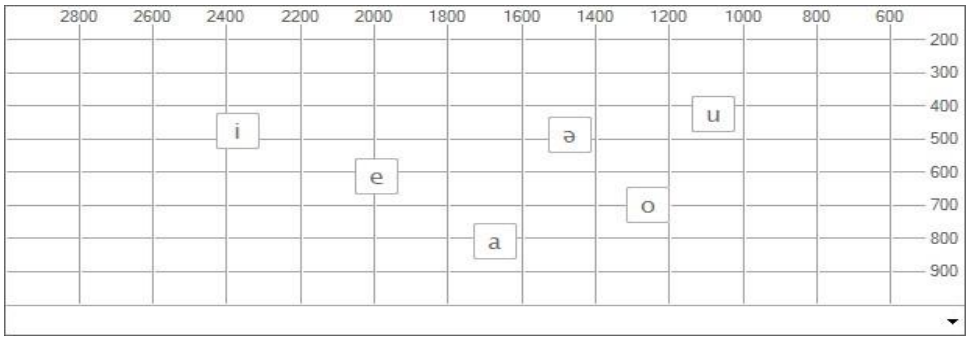


Фигура 6. (Чочева Н. 2014)

Изследването продължава със съставяне на произносителен модел на французойка и българка (Фигура 7, Фигура 8).



Фигура 7. Произносителен модел на французойка (Чочева Н. 2014)



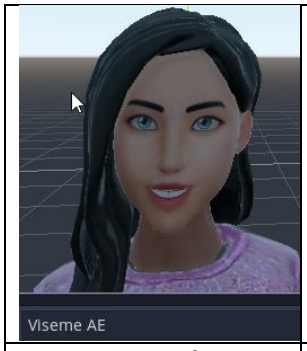
Фигура 8. Произносителен модел на българка (Чочева Н. 2014)

При сравнението на двата модела се забелязва сходство, но ако се използва формантно разпознаване, за получаване на по-точни резултати са необходими допълнителни настройки спрямо избрания език.

Недостатък при този експериментален подход се явява използването само на гласни, а при формиране на висемните форми участие вземат и съгласните. Това налага провеждане на допълнително изследване спрямо възможността за разпознаване на съгласни. Като частично решение в експеримента се използва информацията за съгласните, взета от субтитрите в диалога в комбинация с разпознаването на гласните посредством форманти. За целта е направена синхронизация по време между възпроизвеждането на звуковия файл и изобразяването на букви от субтитрите, но резултатите не могат да бъдат съпоставими спрямо тези с предварително анализиране на речта. Решения от такъв тип все пак могат да се използват в непретенциозни игрови приложения.

На този етап от разработка е създаден речник от формантите от таблица 1 и към звука от възпроизвеждания аудио файл се прилагат по два филтъра за всеки формант в реално време. Резултатът, при който се получава максимален филтриран сигнал се приема за разпознат формант. Във втори речник е зададено съответствие на формантите с висемните форми на устата за използвания перонажен 3D модел. Моделите, създадени с уеб услугата Ready Player Me са с вградени висемни форми (Таблица 2) и не се налага да се анимират допълнително.

При модели, създадени с Make Human или Mixamo/Adobe Character Creator няма предварително зададени висемни анимационни форми и ако са необходими, те трябва да се анимират допълнително.

			
Viseme AE	Viseme Aa	Viseme Ch	Viseme Dd
Viseme E	Viseme Ff	Viseme I	Viseme Kk

Viseme Nn	Viseme O	Viseme PP	Viseme Rr
Viseime Ss	Viseme Sil	Videme Th	Viseme U

Таблица 2

Резултатите от наблюденията на трите описани решения в този раздел водят до идеята за провеждане на нови експерименти, базирани на невронни мрежи, след обучението на които разпознаването на фонемите в аудиофайла и прилагането им към висемни форми ще бъде възможно в реално време без изисквания за допълнителна обработка на резултатите.

Литература

Catford, J.C. (1988): A Practical Introduction to Phonetics, Oxford University Press, стр. 161. ISBN 978-0198242178.

Formanti: https://en.wikipedia.org/wiki/Formant

Jeans, J.H. (1938) Science & Music, reprinted by Dover, 1968.

Manzur, Marque (2018): Manzur A., Marque G., *Godot Engine Game Development in 24 Hours, SamsTeach Yourself, 2018.*

McKendrick, J. G. (1903): Experimental phonetics. In Annual report of the board of regents of the Smithsonian institution for the year ending June 30, 1902 (pp. 241–259). Smithsonian Institution.

Standards Secretariat, Acoustical Society of America, (1994): ANSI S1.1-1994 (R2004) American National Standard Acoustical Terminology, (12.41) Acoustical Society of America, Melville, NY.

Titze, I.R. (1994): Principles of Voice Production, Prentice Hall, ISBN 978-0-13-717893-3.

Titze, I.R. et al (2015): Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization. J. Acoust. Soc. America. 137, 3005–3007.

Чочева Н. 2014: Акустичен анализ на вариативността на гласните във френския и в българския език. Модели за чуждоезиковото обучение. Автореферат на дисертационен труд за присъждане на образователната и научна степен „доктор“. Пловдив 2014.