РОЗРОБКА КОМП'ЮТЕРНИХ ТЕХНОЛОГІЙ МОДЕЛЮВАННЯ ТА КЕРУВАННЯ ВІЗУАЛЬНИМИ ОБРАЗАМИ ЛЮДСЬКОГО ОБЛИЧЧЯ ПРИ СИНТЕЗІ МОВЛЕННЯ[1]

Юрій Крак, Тарас Вінцюк, Микола Кириченко, Федір Гаращенко, Олександр Бармак

Київський національний університет імені Тараса Шевченка
64 вул. Володимирська, Київ 03022
Електронна пошта: krak@unicyb.kiev.ua, vintsiuk@uasoiro.org.ua, kir@dept115.icyb.kiev.ua, garash@unicyb.kiev.ua, barmak@svitonline.com

Yuriy Krak, Taras Vintsyuk, Mykola Kirichenko, Fedir Garashchenko, Olexander Barmak, **Development of a computer technologies for modeling and control of visual images of human face under speech synthesis.** Technologies for modeling of processes animation of arbitrary text of are proposed. Mathematical model of human head with possibilities of facial expression during conversation and synchronization with speech are created.

Вступ

Розвиток сучасних комп'ютерних технологій моделювання і обробки аудіо і відео інформації спрямований на створення систем візуалізації процесу промовляння за допомогою комп'ютерних засобів [1-8]. Важливість використання таких технологій підтверджується їх включенням в MPEG-4 стандарт [9]. Дані технології мають значні прикладні застосування в інтелектуалізації роботи з комп'ютером, кінематографії, телебаченні, телефонії, передачі інформації, тощо. Схема системи озвучення текстів з моделюванням голови людини наведена на рис. 1.

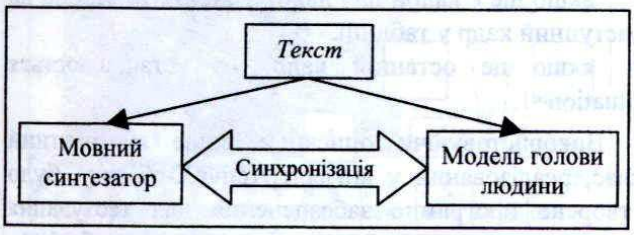

Рис.1.

Для реалізації такої системи потрібно створити програмне забезпечення для побудови об'ємної моделі голови людини і алгоритми, що реалізують динамічну генерацію візуальних образів людського обличчя при синтезі мовлення. Тобто потрібно вміти моделювати процес зміни образів обличчя людини, який синхронний звукам (фонемам), генерованим мовним синтезатором. Припускається, що є мовний синтезатор[10,11], який вхідний орфографічний текст перетворює у розмічений фонетичний текст (транскрипцію), тобто набір фонем з тривалістю кожної фонеми та генерує звук.

В даній роботі пропонується візуалізація процесу промовляння двома методами:

1. Як послідовність змін кадрів зображення конкретного людського обличчя, яке промовляє фонетично розмічений текст – 2D-технологія.
2. Як анімацію (плавний перехід від однієї моделі до іншої) послідовностей об'ємних 3D-моделей людського обличчя, яке промовляє фонетично розмічений текст – 3D-технологія.

Дослідження цих методів, не дивлячись на їх принципову відмінність, показали подібність їх алгоритмічної реалізації. Ця подібність спонукала до створення абстрактного класу, у якому реалізовано перший метод. Важливо відмітити, що після побудови та тестування об'єкта, створеного на основі цього абстрактного класу перехід до другого методу полягає лише у переписуванні відповідних методів об'єкту-нащадку [12-14]. Розглянемо детально два запропонованих методи.

1. 2D-технологія

Алгоритм реалізації першого методу можна описати наступною послідовністю кроків:

1. За допомогою відеокамери знімається людське обличчя, яке промовляє довільний текст. При цьому дотримуються наступні обмеження:
   - 1.1. Камера закріплена на штативі, та сфокусована на голову.
   - 1.2. Фон зйомки має бути однорідним.
   - 1.3. Текст має промовлятися у одному (середньому) темпі.
   - 1.4. При промовлянні тексту голова не повинна рухатися.
   - 1.5. Текст має містити набір фонем за допомогою яких можливо збудувати довільний інший текст (так звана навчальна вибірка).
2. Отримане у п.1 аналогове зображення з допомогою стандартних засобів перетворюється у цифрове. При цьому дотримуються наступні обмеження:
   - 2.1. Має бути файл формату AVI.
   - 2.2. Файл має містити 30 кадрів на одну секунду.
   - 2.3. Розмір кадрів 320x240.

[1] **Робота виконана в рамках ДНТП України "Образний комп'ютер"**

3. З отриманим у п.2 AVI-файлом робляться наступні дії:

3.1. За допомогою програми відео монтажу (ADOBE PREMIER, AVI Constructor, тощо) вибираються з AVI-файлу послідовності кадрів на яких зображено процес промовляння конкретних фонем.

3.2. Вибрані послідовності кадрів запам'ятовуються у BMP-файлах.

3.3. Послідовності BMP-файлів, на яких зображене промовляння конкретної фонеми (трифону), записуються у директорії, назви яких містять назви трифонів.

4. Інформація, створена у п.3, переноситься, за допомогою спеціального програмного забезпечення у реляційну базу даних. У цій базі даних кожній фонемі (трифону) співставленні послідовності кадрів, на яких зображено процес промовляння. База даних (реалізована у ***Paradox***) складається з двох таблиць (таблиці по полю ***Id_Phonemes*** зв'язані як ***Master → Detail***) (Рис. 2).

5. Використовуючи створену у п.4 базу даних, моделюється процес промовляння, як послідовність фонем (трифонів) із транскрипції та відповідні їм послідовності кадрів (із бази даних), які міняються відповідно до тривалостей звучання конкретних фонем.

6. Описаним у пунктах 1-4 способом створюються бази даних з різними дикторами. Користувач має можливість, вказуючи шлях до конкретної бази даних, вибирати диктора для промовляння тексту.

Таблиця ***Phonemes*** (***Master***-таблиця):

№	*Field Name*	*Type*	*Size*	*Key*
1	Id_Phonemes	+		*
2	Phoneme	A	1	***Secondary index***
3	BeforePhoneme	A	1	
4	AfterPhoneme	A	1	
5	Duration	I		

Таблиця ***Pictures*** (***Detail***-таблиця):

№	*Field Name*	*Type*	*Size*	*Key*
1.	Id_Picture	+		*
2.	Id_Phoneme	I		
3.	Item	I		
4.	Picture	G		

Рис. 2

Використовуючи парадигму об'єктно-орієнтованого проектування [13], створюється абстрактний клас TTalkingFace, який і буде моделювати процес промовляння (п.5). До основних полів класу відносяться поля для роботи з базою даних [15-17]:

- база даних (DB);
- джерела даних (DataSourse);
- реляційні таблиці (Table);
- візуальний проглядач графічної інформації (DBImage) .

Основною подією, за якою працюватиме алгоритм візуалізації буде подія від таймеру, тобто подія, яка сигналізуватиме про закінчення заданого інтервалу часу (тривалість звучання конкретної фонеми).

Опишемо основні методи абстрактного класу TtalkingFace.

Метод TimerTimer полягає в наступному:

- Вимикається таймер.
- Запам'ятовується початковий час. В залежності від значення поля situation (1 чи 2) викликаються, відповідно методи Situation1 або Situation2.
- Запам'ятовується кінцевий час.
- Розраховується інтервал для "спання" таймеру на основі витраченого часу (кінцевий час мінус початковий час).
- Вмикається таймер.

Метод Situation1 реалізує алгоритм:

- збільшується лічильник поточної фонеми.
- якщо дійшли до останньої фонеми, то – кінець.
- будується трифон (до поточної фонеми приєднується попередня та наступна фонеми).
- шукається в базі даних рядок з поточним трифоном.
- визначається кількість кадрів для візуалізації фонеми (із бази даних).
- відповідно до вхідної тривалості фонеми та кількості кадрів – розраховується керуючий вектор тривалостей візуалізації кадрів.
- Встановлюється situation=2.

Метод Situation2: використовуючи керуючий вектор тривалостей візуалізації кадрів:

- якщо ще є кадри для даної фонеми, то стаємо на наступний кадр у таблиці.
- якщо це останній кадр, то встановлюється situation=1.

Використовуючи описаний вище абстрактний клас, реалізований у вигляді ActiveX-об'єкту, було створене програмне забезпечення для тестування запропонованого алгоритму. Для тестування бралось моделюватися промовляння фрази: "Добрий день, Україно!". Для створення бази даних, на відео була записана навчальна вибірка фраз, з яких можна було б вибирати фонеми для моделювання. Вибірка складалася з наступних слів:

Слово для відеозапису	*Трифон для бази даних*		
	перед	*фонема*	*після*
Дорога	#	д	о
Здоба	д	о	б
Обрій	о	б	р
...			
Гаї	а	й	і
Кіно	і	н	о
Кіно	н	о	#

Трифони, з відповідними відеокадрами промовляння фонем, були занесені у базу даних

(Рис.3). У відповідній програмі (з реалізованим об'єктом TalkingFace (Рис.4), був змодельований процес візуалізації промовляння фрази "Добрий день, Україно!".

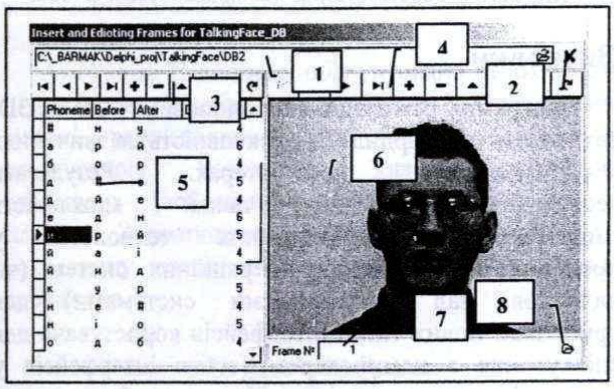

Рис. 3

1. *Вибір бази даних з образами дикторів;*
2. *Вихід з програми;*
3. *Навігатор для роботи з таблицею* ***Phonemes;***
4. *Навігатор для роботи з таблицею* ***Pictures;***
5. *Таблиця* ***Phonemes;***
6. *Поле* ***Picture*** *із таблиці* ***Pictures;***
7. *Поле* ***Item*** *із таблиці* ***Pictures;***
8. *Виведений на кнопку метод* ***LoadFromFile*** *для завантаження кадру з BMP-файлу у поле* ***Picture*** *таблиці* ***Pictures;***

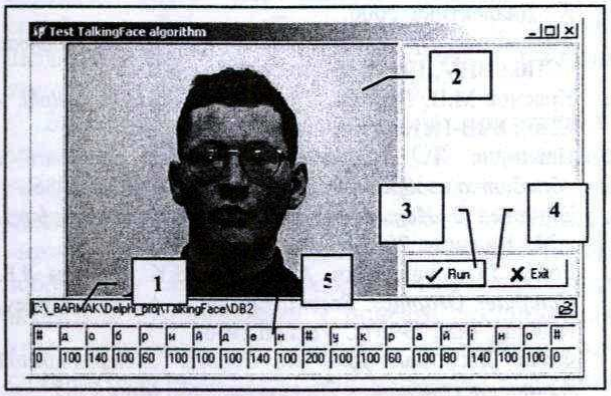

Рис.4

1. *Вибір шляху до бази даних з записами фонем та візем.*
2. *ActiveX об'єкт* ***TalkingFace.***
3. *Запуск візуалізації промовляння фрази з тривалостями.*
4. *Завершення роботи з програмою.*
5. *Панель для завдання фраз та тривалостей фонем.*

2. 3D-технологія

Використовуючи отримані вище результати, створимо технологію та алгоритмічну реалізацію візуалізації промовляння на основі 3D-моделей. Створивши абстрактний клас, у якому реалізований 2D спосіб, та, на його основі, об'єкт, розроблено метод, за допомогою якого перехід до 3D способу буде полягати лише у переписуванні відповідних методів об'єкту-нащадку. Виходячи з цього розглянемо дії, необхідні для реалізації методів 3D-анімації візуалізації промовляння:

1. У пакеті програм трьохмірного моделювання 3D Studio MAX [2] створимо 3D-модель людської голови з морферами, які дозволятимуть керувати мімікою обличчя:
   - 1.1. Беручи за зразок кадри з бази даних на яких зображено процес промовляння трифонів, методом клонування та морфінгу створюється набір 3D-моделей, які їм відповідають.
   - 1.2. Нанесемо на отриманні моделі у якості текстур відповідні їм кадри з бази даних 2D-технології.
   - 1.3. Зробимо експорт отриманих моделей у файли з форматом ASE (ASCII scene export).
   - 1.4. Послідовності ASE-файлів та відповідних їм BMP-файлів (з текстурами) запишемо у директорії, назви яких містять назви трифонів.

2. На основі об'єктів програми 2D-технології розроблено нове програмне забезпечення для роботи з базою даних, яка зберігатиме трифони та відповідні їм моделі. Інформація, створена у попередньому пункті (1.4), переноситься, з допомогою цього програмного забезпечення у реляційну базу даних. У цій базі даних кожній фонемі (трифону) співставленні послідовності 3D-моделей, які зображають процес промовляння. На відміну від бази даних для 2D-технології, у новій базі даних будуть зберігатися не графічні образи, а інформація для інтерактивної побудови 3D-моделі: масив координат вершин трикутників, масив нормалей у кожній вершині, масив текстурних координат для кожної вершини трикутника, масив пікселів самої текстури. Структура таблиці ***Models*** (яка відповідає таблиці ***Pictures*** у 2D-технології) представлена на рис. 5.

Таблиця ***Models*** (***Detail***-таблиця):

№	*Field Name*	*Type*	*Size*	*Key*
1	Id_Model	+		*
2	Id_Phoneme	I		
3	Item	I		
4	Count_Triangles	I		
5	Vertexes	B		
6	Normals	B		
7	Textures	B		
8	Pictures	B		

Рис. 5

3. Використовуючи створену у п.2 базу даних та абстрактний клас TTalkingFace моделюється процес промовляння, як послідовність фонем (трифонів) із транскрипції та відповідні їм послідовності моделей (із бази даних), які міняються відповідно до тривалостей звучання конкретних фонем. У об'єкті, створеному на базі абстрактного класу TTalkingFace

переписується метод роботи з базою даних. Замість візуалізації кадру по події наступного запису у базі даних, створюється метод, який буде по цій події робити рендерінг 3D-моделі, тобто 2D-візуалізацію абстрактної сцени, існуючої у вигляді масивів вершин, масивів нормалей у цих вершинах та масиву текстурних координат.

4. Описаним у пунктах 1-2 способом створюються бази даних з різними моделями-дикторами. Користувач має можливість, вказуючи шлях до конкретної бази даних, вибирати певного диктора для промовляння тексту.

Використовуючи абстрактний клас TTalkingFace з методом для рендерінгу 3D-моделей було розроблене програмне забезпечення для реалізації промовляння по 3D-технології (п. 3). Для моделювання бралася таж сама тестова фраза "Добрий день, Україно!", що і в 2D-технології. Була створена база даних з трифонами і відповідними 3D-моделями для можливості промовляння цієї фрази (Рис.6).

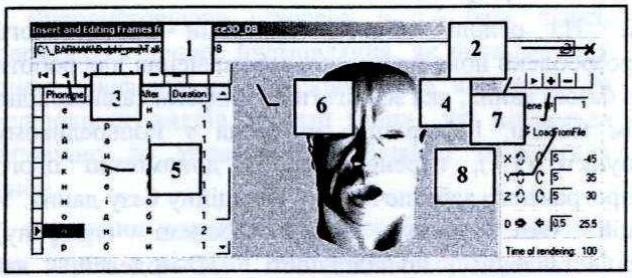

Рис. 6

1. *Вибір бази даних з образами дикторів;*
2. *Вихід з програми;*
3. *Навігатор для роботи з таблицею* ***Phonemes;***
4. *Навігатор для роботи з таблицею* ***Models;***
5. *Таблиця* ***Phonemes;***
6. *3D-двигун;*
7. *Поле* ***Item*** *із таблиці* ***Models;***
8. *Виведений на кнопку метод* ***LoadFromFile*** *для завантаження моделі з ASC-файлу.*

У відповідній програмі був змодельований процес візуалізації промовляння фрази "Добрий день, Україно!" (див.Рис.7).

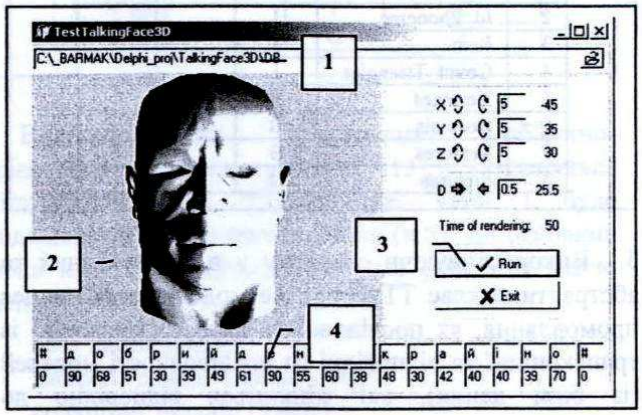

Рис. 7

1. *Вибір шляху до бази даних з записами моделей.*
2. *Панель для рендерінгу.*
3. *Запуск візуалізації промовляння фрази з тривалостями*
4. *Панель для завдання фраз та тривалостей фонем.*

Висновки

Програмна реалізація запропонованих 2D та 3D-технологій підтвердила їх ефективність на звичайних мультимедійних комп'ютерах. Результати тестування показали також можливість імплементації запропонованих технологій у програмне забезпечення операційних систем (чи надбудов над операційними системами) для організації нових типів інтерфейсів користувача для спілкування з комп'ютером (тобто інтерфейсів у яких спілкування користувача з комп'ютером та комп'ютера з користувачем відбувається з допомогою звичайної мови). Подальший розвиток таких технологій полягає в розробці нових математичних методів для побудови об'ємних зображень голови людини і створення якісних синтезаторів української мови.

Література

1. Флеминг Б., Доббс Д. *Методы анимации лица. Мимика и артикуляция*. Пер. с англ. – М.: ДКМ Пресс, 2002 .
2. Мердок К. *3D Studio MAX R3. Библия пользователя.* – К.: Диалектика, 2000.
3. Тихомиров. *Программирование трехмерной графики.* –СПб.: BHV, 1998
4. Краснов М.В. *OpenGL. Графика в проектах Delphi.* – СПб.: БЧВ-Петербург, 2001.
5. Павлидис Т. *Алгоритмы машинной графики и обработка изображений.* – М.: Радио и связь, 1986.
6. Эйнджел Э. *Интерактивная компьютерная графика.* – М.: Вильямс, 2001.
7. Foley J.D., van Dam A., Feiner S.K., Hughes J.F. *Computer Graphics, Second Edition.* – Addison-Wesley, Reading, MA, 1990 (C Version 1996).
8. Фоли Дж., ван Дэм А. *Основы интерактивной машинной графики: в 2-х книгах.* – М.: Мир, 1985
9. Pandzic I.S., Ostermann J., Millen D. *User evaluation: Synthetic talking faces for interactive services*. The Visual Computer. 15, 1999. – pp. 330-340.
10. Т.К. Винцюк. *Анализ, распознавание и смысловая интерпретация речевых сигналов.* – Киев: Наукова думка, 1987.
11. Дж.Л.Фланаган. *Анализ, синтез и восприятие речи.* Пер. с англ. М.:Связь, 1968
12. Каханер Д., Моулер К., Нэш С. *Численные методы и программное обеспечение.* – М.: Мир, 2001
13. Г.Буч. *Объектно-ориентированный анализ и проектирование.* – М., Бином, 1998, 558 с.
14. Миллер Т., Пауэл Д. *Использование DELPHI 3.* – К.: Диалектика, 1997.
15. Дж.Мартин. *Организация баз данных в вычислительных системах.* – М., Мир, 1980. 662 с.
16. К.Дейт. *Введение в системы баз данных.* – М., Наука, 1980.
17. Д.Мейер. *Теория реляционных баз данных.* - М., Мир, 1987. 608 с.