УДК 004.928

М. Давидов
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

СИНТЕЗ ВИДИМОЇ АРТИКУЛЯЦІЇ ВІРТУАЛЬНОГО ПЕРСОНАЖА З АУДІОПОТОКУ ДЛЯ СИСТЕМИ СУРДОПЕРЕКЛАДУ

Описано метод синтезу видимої артикуляції віртуального персонажа з аудіопотоку для системи сурдоперекладу. На відміну від методів, які працюють із записом цілого речення для синтезу видимої артикуляції із затримкою в часі, запропонований метод працює із наперед заданою затримкою, яка не залежить від довжини речення. Метод оснований на виборі кадрів навчального відео в різній послідовності для забезпечення максимально реалістичної артикуляції.

Ключові слова: анімація губ з аудіопотоку, видима артикуляція, сурдопереклад.

Lips animation method for synchronous translation into sign language is described. As opposed to methods that work with a record of a full sentence, the proposed solution has a constant delay that does not depend of sentence length. The method is based on selection of video frames from a training video in a specific order to achieve natural look.

Key words: speech driven lip animation, visible articulation, sign language translation.

Вступ. Постановка проблеми

Задача анімації руху губ розглядається в контексті побудови системи сурдоперекладу реального часу. Побудова системи автоматичного перекладу словесної мови на жестову є актуальною проблемою, вирішення якої розширить можливості спілкування для людей із вадами слуху. Такий переклад може здійснюватися з тексту, аудіозапису або аудіопотоку. Роботи зі створення систем сурдоперекладу вже здавна ведуться за кордоном. Наприклад, відома розробка iCommunicator [1] дає змогу перекладати аудіозапис речення на текст англійською мовою, перекладати речення мовою жестів та озвучувати текст речення. Недоліком відомих систем є відсутність можливості перекладу із малою часовою затримкою та підтримка лише калькованої жестової мови.

Проблеми сприймання голосової інформації людьми із пониженим слухом виникають у багатьох життєвих ситуаціях, наприклад, у разі спілкування з людиною, яка не володіє мовою жестів, у телефонній розмові, у відеотелефонній розмові, під час прослуховування голосових повідомлень на вокзалах, аеропортах, проглядання телепередач, які йдуть без субтитрів або сурдоперекладу.

Щоб полегшити сприймання голосової інформації людьми із пониженим слухом, інформація може доповнюватись текстовими повідомленнями (субтитрами), видимою артикуляцією, мануальною мовою [2] та жестовою мовою (сурдоперекладом).

Як показали дослідження, які провів К. Саммерфілд (Q. Summerfield) [3], видима артикуляція займає важливе місце в процесі сприймання голосової інформації людьми із послабленим слухом. Точність сприймання інформації людьми із послабленим слухом при читанні з губ співрозмовника виявилася на 43 % вищою, ніж у разі прослуховування лише голосу. Неможливість слідкувати за губами співрозмовника спричиняє істотні незручності під час телефонної розмови та значно ускладнює процес спілкування. Більшість систем відеотелефонії поки що не вирішують проблеми відтворення артикуляції, оскільки швидкість передавання відеосигналу в таких мережах недостатня для чіткого відтворення руху губ співрозмовника.

У лабораторії жестової мови Львівської політехніки ведуться роботи з метою створення системи сурдоперекладу, яка перекладає речення із мінімально можливою затримкою. Автоматична система сурдоперекладу, яка розробляється, складається з програмних модулів розпізнавання

мовлення, перекладу жестовою мовою, синтезу жестикулювання, синтезу артикуляції та формування відео (рис. 1). Одним із важливих елементів системи сурдоперекладу є модуль синтезу артикуляції, який дає змогу людині, яка не чує, читати по губах для ліпшого розуміння повідомлення.

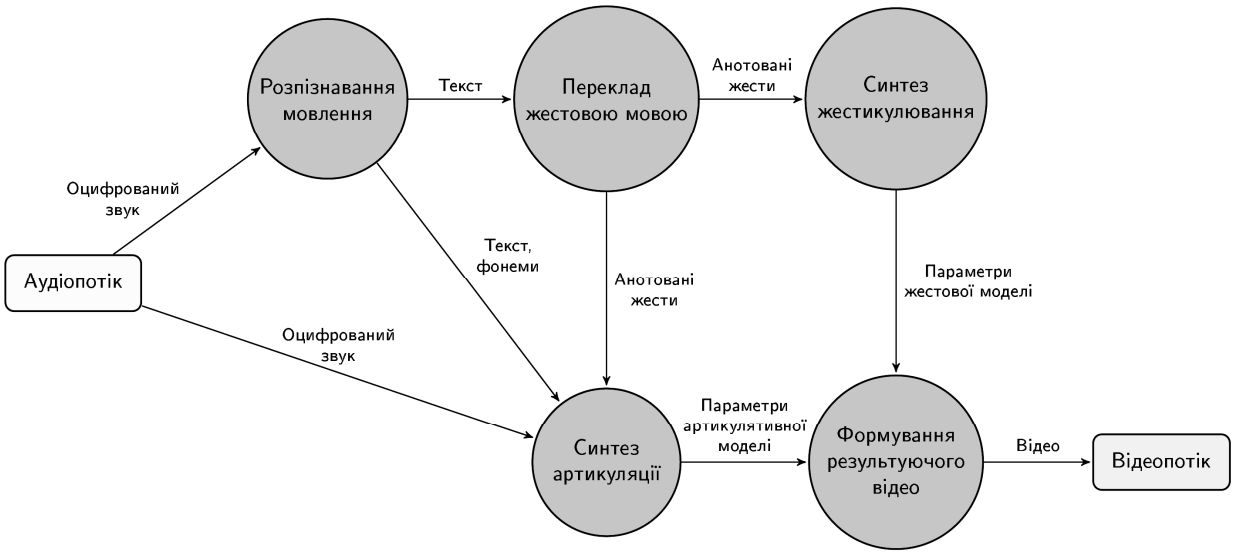

Рис. 1. Програмні модулі та потоки даних у системі сурдоперекладу

У системі сурдоперекладу анімація губ може виконуватися в таких випадках:

1) анімація губ за аудіопотоком для полегшення сприйняття звукової інформації людьми із частковою втратою слуху;

2) анімація губ за аудіопотоком у випадку, коли модуль розпізнавання мовлення не може правильно розпізнати слова;

3) анімація губ при дактилюванні слова;

4) анімація губ при жестикулюванні.

Сьогодні найдослідженіший третій та четвертий випадки анімації губ – анімація під час жестикулювання та дактилювання [4]. Основна увага у статті зосереджена на анімації губ для полегшення сприйняття звукової інформації людьми з частковою втратою слуху та анімації у разі відсутності фонемної інформації.

Аналіз останніх досліджень та публікацій

Через дуже різноманітну міміку обличчя під час розмови та чутливість людини до тонкощів міміки створення реалістичного персонажу, що розмовляє, залишається складною задачею комп’ютерної графіки. Відомі два основні підходи для синтезу видимої артикуляції віртуального персонажа [5]. Перший оснований на виділенні фонем із вхідного тексту або звукозапису речення. При цьому створення відео відбувається лише за фонемною інформацією. Другий підхід оснований на аналізі вхідного звукового сигналу без виділення фонем.

Перший підхід може застосовуватися як для синтезу видимої артикуляції із тексту речення, так і зі звукозапису. Із використанням словника цей підхід дає змогу розпізнати близькі за звучанням звуки, які мають різну видиму артикуляцію, і синтезувати правильну анімацію. Проте використання словника узалежнює від обмеженого набору слів, а слова, які погано розпізнані, можуть ускладнити розуміння повідомлення. З переходом до фонемного подання також втрачається емоційна складова речення, сміх, інтонація. Сучасні засоби автоматичного синтезу мови з тексту речення не здатні правильно розставити логічні наголоси і тому обмежено застосовуються для створення реалістичних анімацій. Використання систем зі словником викликає великі затримки в системах реального часу, оскільки система чекає вимови повного слова або речення, а потім синтезує видиму артикуляцію.

Системи, які синтезують видиму артикуляцію на основі характеристик звукового сигналу, дають змогу відтворити емоційну складову речення, логічний наголос, сміх, позіхання. Анімація може зображати звуки інших мов, навіть якщо вони не містяться в навчальному наборі. Перенесення таких систем на іншу мову полягає лише у створенні навчального відео іншою мовою.

Синтез анімації на основі звукозапису має і недоліки. Наприклад, складно розпізнати деякі звуки, які дуже близькі за звуковими характеристиками, але сильно відрізняються артикуляцією вимови. Такими є, наприклад, звуки /м/ та /н/, /б/ та /в/.

Для синтезу відео на основі звукового потоку відомі закордонні розробки використовують метод прихованих марковських моделей та його модифікації, метод Калмана, метод нейронних мереж та метод тривізем.

Метод прихованих марковських моделей використали Симон та Кокс (Simons and Cox [6]) для створення синтезованої голови, яка анімується на основі звукозапису мови. Для створення навчальної бази опрацьовано 50 речень, супроводжених відео. З навчальної бази видобуто 10000 векторів, що характеризують звук, та 5000 векторів, що характеризують зображення обличчя. За допомогою дискретизації векторів створено словник із 16 символів для зображення обличчя та 64 символів для опису звуку. Створено повнозв'язну марковську модель із 16 станами. Кожен стан відповідав стану обличчя та в кожному стані синтезувався один кадр обличчя. Модель Маркова була навчена за допомогою алгоритму Баума–Уелша.

Метод фільтрації Калмана, використаний у роботі [7], показав краще згладжування анімації губ, ніж анімація, синтезована із використанням прихованих марковських моделей. Незважаючи на те, що ця анімація мала менше рваних рухів, вона, за даними авторів, більше відрізнялася від справжньої.

У системі 'Picture my voice' [8] використана рекурентна нейронна мережа, яка перетворює з простору 11x13 коефіцієнтів косинусного перетворення Фур'є на 37 параметрів анімації обличчя. Навчальні дані одержані із системи анімації обличчя на основі фонем. Під час синтезу анімації фонемне подання не використовується.

Метод коартикуляції, застосований у [9], використовує проміжний фонемний запис речення. На етапі навчання визначаються траєкторії руху контрольних точок обличчя під час вимови можливих трійок та пар фонем. На етапі синтезу за алгоритмом динамічного програмування визначається оптимальна траєкторія руху контрольних точок обличчя. Метод показує найкращу якість синтезованого відео для англійської мови. Недоліком методу є необхідність створювати велику навчальну базу та неможливість синтезувати синхронне відео реального часу.

Формулювання цілі статті

Основна увага статті зосереджена на синтезі видимої артикуляції, коли анімація виконується у реальному часі безпосередньо з аудіопотоку без виділення фонемної інформації. При цьому задається параметр максимальної затримки ΔT синтезованої анімації відносно вхідного звукового потоку.

Основна частина

Розроблений модуль синтезу артикуляції складається з блока виділення характеристик аудіопотоку та блоку синтезу візем (рис. 2).

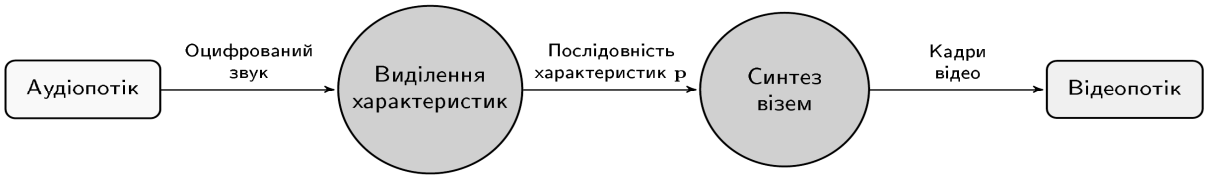

Рис. 2. Структура програмного модуля синтезу артикуляції

Оцифрований звук надходить із мікрофона або мережі у вигляді послідовності відліків $\{x(i)\}$ із частотою дискретизації звуку F. Блок виділення характеристик звукового сигналу опрацьовує послідовність $\{x(i)\}$ з метою виділення її локальних характеристик. Через те, що послідовність значень $\{x(i)\}$ не є показовою характеристикою звукового сигналу [10, с. 614], використано одну із поширених характеристик звукового потоку – енергію звуку на різних частотах та її зміну в часі.

Для вибору характеристик аудіопотоку звуковий потік розглядається як послідовність відліків $\{x(i)\}$, із частотою дискретизації $F = 8кГц$. Звуковий потік $x_1(i)$ з іншою частотою дискретизації F_1 перетворюється до частоти дискретизації F лінійною інтерполяцією:

$$x(i) = x_1\left(\left\lfloor i\frac{F_1}{F}\right\rfloor\right)\left(i\frac{F_1}{F} - i\frac{F_1}{F}\right) + x_1\left(i\frac{F_1}{F}\right)\left(1 - i\frac{F_1}{F} + i\frac{F_1}{F}\right).$$

Для виділення енергії звуку в околі відліку q на частоті f використано згортку з фільтром

$$E(q,f) = \sqrt{\left(\sum_{i=q-s}^{q+s} x(i)\sin\left(2\pi\frac{f}{F}i\right)e^{-\alpha^2\left(\frac{i-q}{F}\right)^2}\right)^2 + \left(\sum_{i=q-s}^{q+s} x(i)\cos\left(2\pi\frac{f}{F}i\right)e^{-\alpha^2\left(\frac{i-q}{F}\right)^2}\right)^2},$$

де s – половина розміру вікна, α – параметр функції вікна Лапласа (α – величина близько 400, запропонована в [11]).

Енергія обчислювалася для частот у діапазоні 80…800 Гц із кроком 40 Гц:

$$E_i(q) = E(q, 40i + 40),\ i = 1,...,19.$$

Отримані 19 значень енергії на частотах 80, 120, …, 800 Гц додатково приведено до 8 значень за допомогою зваженого підсумовування:

$$E_i(q) = \sum_{j=(2i+1)-2}^{(2i+1)+2} E_j(q)e^{-\left(\frac{j-(2i+1)}{4}\right)^2},\ i = 1,...,8,$$

де $E_i(q)$ – зважена сума енергії сигналу на частотах $80i...80i+160Гц$.

Для запобігання шуму та для визначення зміни енергії в часі обчислення енергії проведено в чотирьох точках із кроком 128 відліків.

$$\overline{E}_i(q) = \frac{E_i(q-192) + E_i(q-64) + E_i(q+64) + E_i(q+192)}{4},\ i = 1,...,8,$$

$$D_i(q) = \frac{-3E_i(q-192) - E_i(q-64) + E_i(q+64) + 3E_i(q+192)}{4},\ i = 1,...,8,$$

де $\overline{E}_i(q)$ – середнє значення енергії в околі точки q на частотах $80i...80i+160Гц$, $D_i(q)$ – середнє значення швидкості зміни енергії в околі точки q на частотах $80i...80i+160Гц$.

Для опису характеристик звукового сигналу в часовому околі q використано 16 значень $<\overline{E}_1(q),...,\overline{E}_8(q),D_1(q),...,D_8(q)>$. Характеристики сигналу визначено з кроком $8000/FPS$, де FPS – частота кадрів на секунду.

На виході блока виділення характеристик отримаємо послідовність $\mathbf{p} = p_1p_2...p_N$ векторів $p_i \in P$, які описують характеристики аудіопотоку через рівні проміжки часу $1/F_v$, де F_v – частота кадрів відео, яке буде створюватися. Кількість відліків N, які доступні під час синтезування наступного кадру відео, задається параметром максимальної затримки синтезу ΔT:

$$N = \Delta T \cdot F_v.$$

Послідовність $\mathbf{p}$ надходить на блок аудіо-відео перетворення для синтезу відео. На виході блока отримаємо послідовність $\mathbf{v} = v_1v_2...v_i...$ векторів $v_i \in V$, які однозначно описують кожен кадр

відео, яке буде створюватися. Блок синтезу відео створює відеопотік та суміщає його зі звуком. На кожному кроці роботи системи синтезується один кадр зображення, після чого одержується наступний елемент послідовності **p**, а перший її елемент видаляється.

Для формування векторів, які описують кадри вихідного відео, використано приховану марковську модель $M = < T, P, A, B >$, де T – стани марковської моделі, P – множина можливих спостережень, $A: T \times T \times P \to \mathbb{R}$ – імовірності переходу між станами, $B: T \times P \to \mathbb{R}$ – імовірності отримання спостережень у станах моделі. На відміну від відомих підходів, імовірності переходу між станами прихованої марковської моделі розраховують залежно від отриманого спостереження, що дало змогу звузити область пошуку оптимальної послідовності станів та одержати природнішу анімацію.

Станами T використаної прихованої марковської моделі є всі кадри навчального відео $q_1 q_2 \ldots q_m$. Позначимо $Q: T \to P$ значення показників аудіопотоку в навчальних кадрах та $W: T \to V$ – вектори, які кодують зображення обличчя в них.

Задача синтезу видимої артикуляції розглядається як задача пошуку за вхідним аудіопотоком послідовності станів $t_1 t_2 \ldots t_n$, $t_i \in T$ так, щоб синтезовану послідовність кадрів на основі послідовності векторів $W(t_1), W(t_2), \ldots, W(t_n)$, які кодують зображення обличчя в кадрі, людина сприймала як природну.

Для цього на кожному кроці вибору наступного кадру виконується пошук послідовності станів прихованої марковської моделі, яка максимізує їхню апостеріорну імовірність за відомим спостереженням

$$\langle t_1, t_2, \ldots, t_N \rangle = \underset{t_1, t_2, \ldots, t_N}{\arg\min} \; P(t_1, t_2, \ldots, t_N \mid t_0, p), \tag{1}$$

де t_0 – стан марковської моделі, який використаний для синтезу попереднього, вже показаного кадру відео; p – послідовність відомих характеристик звукового сигналу. Імовірність послідовності станів у модифікованій марковській моделі розраховують за формулою

$$P(t_1, t_2, \ldots, t_N \mid t_0, p) \sim P(p \mid t_1, t_2, \ldots, t_N) \cdot P(t_1, t_2, \ldots, t_N \mid t_0) = \prod_{i=1}^{N} B(t_i, p_i) \cdot A(t_{i-1}, t_i, p_i). \tag{2}$$

Для розрахунку імовірності виходу марковської моделі використано нормальний розподіл із модою, заданою навчальним відео:

$$B(t_i, p_i) = \exp\left(-k_1 \cdot |Q(t_i) - p_i|^2\right),$$

де k_1 – коефіцієнт впливу характеристик аудіопотоку.

Розрахунок імовірностей переходу виконується із пошуком C кадрів навчального відео, у яких отримано характеристики аудіопотоку, найближчі до поточних. Також визначають один стан, до якого завжди можливий перехід – наступний кадр у навчальному відео. Для множини вибраних станів $\{t_j\}$ імовірність переходу встановлюється пропорційно до імовірності синтезу спостереження та подібності кадрів

$$A(t_{i-1}, t_j, p_i) \sim B(t_j, p_i) \cdot \exp\left(-k_2 |W(t_j) - W(t_c)|^2\right),$$

де k_2 – коефіцієнт, який враховує подібність суміжних кадрів синтезованого відео та зменшує кількість видимих стрибкоподібних змін артикуляції.

На рис. 3 проілюстровано процес вибору можливих станів моделі для продовження синтезу відео. У наведеному прикладі t_0 – останній кадр, який вже синтезовано. Цьому кадру відповідає кадр q_2 навчального відео, який є останнім відомим станом марковської моделі. Для синтезу кадру t_1 можливий перехід до станів q_3, q_5, q_6. При цьому q_3 вибирають як наступний до q_2, а q_5 та q_6 вибирають як такі, характеристики аудіопотоку яких найближчі до p_1. У прикладі $C = 2$.

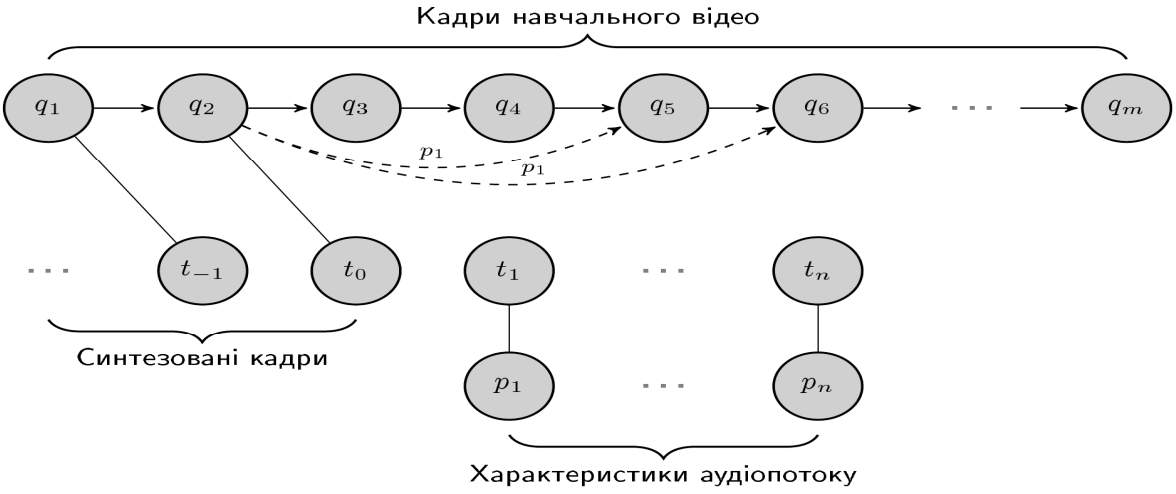

Рис. 3. Можливі варіанти вибору наступного стану моделі

Пошук множини із C станів для продовження виконується із використанням R-дерева пошуку, що забезпечує час пошуку $O(C\log(|T|))$. Максимум виразу (1) розраховують із використанням динамічного програмування. Враховуючи, що на кожному кроці кількість станів зростає не більше, ніж на C, максимальна кількість станів не перевищує $1+TC$. Це забезпечує пошук максимуму за час $O\left(T(1+TC)^2\right)$ або $O\left(T^3C^2\right)$. Для зменшення обчислювальної складності пошуку максимуму вирішено на кожному кроці залишати лише Q найімовірніших станів, що дало змогу зменшити час пошуку до прийнятних $O\left(TQ^2\right)$ без видимого погіршення результату. У проведених експериментах значення T>30 не приводили до поліпшення синтезованої артикуляції, оскільки це більше, ніж середній час вимовляння одного слова. Значення Q вибирали в діапазоні 10...20, що давало змогу застосовувати алгоритм у реальному часі.

Результати експериментів

Для навчання розробленої системи синтезу видимої артикуляції створено відеозапис із віртуальним персонажем, який вимовляє навчальний монолог. Відеозапис створено за допомогою програми Poser із використанням фрази німецькою мовою «Guten Tag lieber Anrufer! Sie sind verbunden mit meiner Video-Mailbox. Zur Zeit bin ich leider nicht persönlich erreichbar. Bitte hinterlassen Sie eine Nachricht nach dem Signalton» довжиною 10 секунд та частотою 30 кадрів на секунду.

На основі навчального відео синтезовано анімацію для аудіопотоків німецькою, англійською та українською мовою. Приклад послідовності кадрів для одного слова синтезованої анімації англійською мовою наведено на рис. 4. Час синтезу одного кадру відео становив 20 мс для $N = 5$ та $C = 10$ на комп'ютері з процесором Intel i5 із частотою 3,1 ГГц.

Рис. 4. Кадри синтезованої послідовності візем при вимові слова "Hello!".
Кадри вибрано через кожні 5 кадрів відеоряду

Створений прототип використовується для синтезу відео на сервері, який обслуговує систему голосового меню для відеотелефонів. Як показали експерименти, користувачам із пониженим слухом легше зрозуміти повідомлення, доповнене синтезованим відео, ніж повідомлення, яке містило лише звук. Через відсутність програм, які синтезують анімацію губ у реальному часі,

порівняння здійснювалося з програмою LipSync компанії Annosoft, яка є одним зі світових лідерів у розробленні засобів синхронізації анімації губ. Програма LipSync виконує попереднє опрацювання цілого файла для виділення фонем. В експериментах із синтезу артикуляції для німецької, англійської та української мови істотних недоліків у синтезованій анімації порівняно з програмою LipSync не виявлено.

Висновки

У ході проведених робіт побудовано систему синтезу видимої артикуляції для допомоги людям із вадами слуху. Віртуальні персонажі, які розмовляють, також можуть застосовуватись на інтерактивних web-сторінках, у системах відеотелефонії, ігрових системах тощо. Подальші дослідження будуть спрямовані на вдосконалення аудіовізуального перетворення, яке зможе анімувати не лише мову, а й інші звуки, такі як дзвінок, крик, сміх тощо.

1. iCommunicator Features and Benefits. [Електронний ресурс]. – Режим доступу: http://www.myicommunicator.com/productinfo/features_benefits.shtml. – перевірено 10.10.2013. 2. Heracleous P. Gestures and Lip Shape Integration for Cued Speech Recognition / P. Heracleous, N. Hagita, D. Beautemps // Pattern Recognition (ICPR), 2010 20th International Conference on. – 23-26 Aug. 2010. – P. 2238–2241. 3. Summerfield Q. Lipreading and audio-visual speech recognition / Q. Summerfield // Phil. Trans. R. Soc. Land. B. – Vol. 335. – 1992. – P. 71–78. 4. Крак Ю.В. Анімація віртуальних образів людського обличчя при синтезі мовлення / Ю.В. Крак, О.В. Бармак // Искусственный интеллект–2002: мат. Международной научно-технической конференции. Таганрог: Изд-во ТРТУ. – Том. 2. – 2002. – С. 138–142. 5. Ezzat T. Trainable videorealistic speech animation / T. Ezzat, G. Geiger, T. Poggio // ACM Transaction on Graphics (Proceedings of ACM SIGGRAPH' 02). – 2002. – P. 388–398. 6. Simons A. Generation of mouth shapes for a synthetic talking head / A. Simons, S. Cox. // Proc. Inst. Accoust. – vol. 12. – 1990. – P. 475–482. 7. Lehn-Schioler T. Mapping from speech to images using continuous state space models / T. Lehn-Schioler, L. K. Hansen, J. Larsen // Book Section. Springer Berlin Heidelberg: Machine Learning for Multimodal Interaction. – Lecture Notes in Computer Science. – 2005. – P. 136–145. 8. Massaro D. W. Picture my voice: Audio to visual speech synthesis using artificial neural networks / D. W. Massaro, J. Beskow, M. M. Cohen, C. L. Fry, and T. Rodriguez // Proc. AVSP 99. – 1999. 9. Deng Z. Synthesizing speech animation by learning compact speech co-articulation models / Z. Deng, J. P. Lewis, and U. Neumann // Proc. of Computer Graphics International (CGI) 2005. – Long Island, NY: IEEE Computer Society Press. – June 2005. 10. Allen J. Natural Language Understanding / James Allen. –2nd ed. The Benjamin Comings Publishing, Inc. – 1995. 11. Сорокин В.Н., Цыплихин А.И. Сегментация и распознавание гласных // Информационные процессы. Том 4. – 2004.– № 2. – С. 202–220.