13. Пинчук, С. И. Диэлектрическая проницаемость оксидов цинка, синтезированных методом жидкофазного химического осаждения [Текст] / С. И. Пинчук, А. А. Внуков, И. Б. Белов, А. С. Баскевич, А. Ю. Ляшков, А. Р. Омельчук // Металознавство та термічна обробка металів. – 2013. – № 4. – С. 48–53.

14. Пинчук, С. И. Влияние технологических параметров химического осаждения на свойства нанокристаллического оксида цинка [Текст] / С. И. Пинчук, А. А. Внуков, И. Б. Белов, А. С. Баскевич, А. Ю. Ляшков, А. Р. Омельчук // Металлургическая и горнорудная промышленность. – 2014. – № 1. – С. 63–65.

15. Горелик, С. С. Рентгенографический и электронно-оптический анализ [Текст] / С. С. Горелик, Ю. А. Скаков, Л. Н. Расторгуев. – М.: МИСИС, 1994. – 328 с.

Рекомендовано до публікації д-р техн. наук, професор Пінчук С. Й.
Дата надходження рукопису 30.10.2017

Внуков Александр Александрович, кандидат технических наук, доцент, кафедра покрытий, композиционных материалов и защиты металлов, Национальная металлургическая академия Украины, пр. Гагарина, 4, г. Днепр, Украина, 49600
E-mail: alvnukov74@gmail.com

Головачев Артем Николаевич, кандидат технических наук, доцент, кафедра электрометаллургии, Национальная металлургическая академия Украины, пр. Гагарина, 4, г. Днепр, Украина, 49600
E-mail: golartem@ukr.net

Белая Алена Викторовна, кандидат технических наук, кафедра покрытий, композиционных материалов и защиты металлов, Национальная металлургическая академия Украины, пр. Гагарина, 4, г. Днепр, Украина, 49600
E-mail: alena@ukr.net

УДК 621.391
DOI: 10.15587/2313-8416.2017.118212

РОЗРОБКА АУДІОВІЗУАЛЬНОЇ СИСТЕМИ РОЗПІЗНАВАННЯ МОВИ

© О. М. Горносталь, Я. Ю. Дорогий

Запропонована модель аудіовізуальної системи на базі прихованих Марківських моделей, яка дозволяє розпізнавати мову в реальному часі. Модель дає інструментарій розпізнавання мови, який можна використати в умовах, де інші засоби можуть бути неможливими, наприклад, в умовах відсутності аудіо складової. Досліджена та перевірена працездатність моделі на прикладі розпізнавання цифр, отримані очікувані результати

Ключові слова: *аудіовізуальна система, приховані Марківські моделі, візема, зв'язані приховані Марківські моделі*

1. Вступ

Існують різні методи розпізнавання мови, проте останнім часом основним став метод порівняння з еталоном. Це пов'язано головним чином з прогресом в області електронних компонентів, зокрема зі збільшенням обчислювальної потужності процесорів і обсягів пам'яті. При зіставленні з еталоном звуки перетворюються в характерні образи, які порівнюються з заздалегідь запасеними еталонними образами, і обчислюється ступінь їхньої подібності. Результатом розпізнавання є найбільш схожий еталонний образ.

При розпізнаванні мови шляхом зіставлення з еталоном виникає кілька проблем, серед яких найбільш типовими є наступні.

1. Тимчасові зміни характерних образів мови. Причиною змін є різна швидкість проголошення одних і тих же звуків, тобто непостійність тривалості звуків. Навіть одні й ті ж слова, вимовлені людиною, кожен раз міняються за тривалістю. Якщо ж одні і ті ж слова вимовляються різними людьми, їхня тривалість може ще більше відрізнятися.

2. Вплив розмірів органу мови на образи. Як вже говорилося вище, розміри органів мови у людей різні. Тому, навіть якщо слова вимовляються органами однакової форми, їх резонансні частоти можуть відрізнятися. На образах це проявляється як індивідуальна особливість людини.

Крім цього існує проблема артикуляційного сполучення, тобто відмінності одного і того ж звуку, зумовлені впливом різних звуків до і після нього, проблема акценту, що виникає за рахунок різниці в манері говорити і в умовах життя й інші проблеми. Для того щоб вирішити проблему артикуляційного сполучення, часто застосовують великі одиниці розпізнавання типу слів, вимовлених з паузою.

Використання візуальних спостережень на додачу до акустичних спостережень в системах автоматичного розпізнавання мови (ASR) зацікавили дослідників як можливе рішення швидкого падіння продуктивності чисто аудіальних ASR систем в зашумлених середовищах. Єдина вимога полягає в тому, що при будь-яких умовах аудіовізуальна (AV) система ASR

повинна розпізнавати, по меншій мірі, не гірше ніж аудіальна. Щоб задовольнити цю вимогу, необхідно динамічно адаптувати внесок кожної модальності в класифікаційні рішення, зроблені аудіовізуальною моделлю. Це досягається шляхом зважування вкладу кожної модальності відповідно до її контенту і надійності, використовуючи так звані потокові ваги (SW).

2. Аналіз літературних даних

З розвитком комп'ютерних систем стає все більш очевидним, що використання цих систем набагато розшириться, якщо стане можливим використання людської мови при роботі безпосередньо з комп'ютером, і зокрема стане можливим управління машиною звичайним голосом в реальному часі, а також введення і виведення інформації у вигляді звичайної людської мови.

Існуючі технології розпізнавання мови не мають доки достатніх можливостей для їх широкого використання, але на даному етапі досліджень проводиться інтенсивний пошук можливостей вживання коротких багатозначних слів (процедур) для полегшення розуміння.

Для успішного розпізнавання мови слід вирішити такі завдання:

– обробку словника (фонемний склад);

– обробку синтаксису;

– скорочення мови (включаючи можливе використання жорстких сценаріїв);

– вибір диктора (включаючи вік, стать, рідну мову і діалект);

– тренування дикторів;

– вибір особливого виду мікрофона (беручи до уваги спрямованість і місце розташування мікрофона);

– умови роботи системи і отримання результату із зазначенням помилок.

Основними математичними засобами для вирішення задачі розпізнавання мови є приховані Марковські моделі (ПММ), нейронні мережі та нечітка логіка.

У моделях з використанням ПММ кожна фонема є чимось на зразок однієї ланки в ланцюзі, з яких складається ціле слово. Під час підстановки різних варіантів фонем, ці ланки можуть змінюватися, утворюючи відразу кілька слів з одного і того ж набору фонем. З цього набору фонем програма намагається побудувати слова. Під час цього процесу програма присуджує кожній фонемі значення ймовірності її вживання в даному контексті.

Далі йде ще більш складний процес формування словосполучень і речень. З цього хаосу фонем програма намагається побудувати логічні ланцюги, з яких в далі виходять цілі речення.

Основними працями із застосуванням ПММ є [1–3]. Рівень розпізнавання, який був досягнутий в цих роботах склав 95 % в умовах відсутності шуму. В праці [4] для попередньої класифікації перед подачею даних на ПММ використана машина опорних векторів (SVM). Досягнута точність розпізнавання на рівні 92 %. Ще одна праця – розпізнавання слів литовської мови [5]. В роботі досягнута точність розпізнавання 80 % на базі з 750 слів.

В якості ознак, які витягнуті з мови, добре відомі LPC (коефіцієнт лінійного передбачення), кепстр, спектр та інші. На спектральному часовому образі (СЧО), за осями якого відкладаються час і частоти, одержувані в результаті поділу мови на короткі інтервали і спектрального аналізу на цих інтервалах, добре виражені особливості мови. Зчитуючи спектр, людина може «читати» по СЧО вимовлені звуки.

Як зазначалося вище, людина вимовляє слова, змінюючи органом мови резонансну частоту, тому особливо важливими в СЧО є резонансні частоти, тобто викиди. Резонансні частоти для голосних звуків називають формант, проте використовують і назву «локальний викид» як розширення поняття форманта на приголосні звуки. У методі розпізнавання сказаного слова, запропонованого в дослідженні [6], розпізнавання здійснюється шляхом визначення, який локальний викид присутній і як він змінюється в часі. Оскільки інтерес представляє лише місце розташування локального викиду, дані можна представити у двійковому вигляді: 1 – на місці локального викиду, 0 – в інших місцях, локалізувавши тим самим положення викиду і скоротивши обсяг даних. Отриманий образ називають двійковим спектральним часовим образом (ДСЧО) і використовують його як особливість мови. Застосування ДСЧО при зіставленні образів полягає в тому, що для слова, вираженого за допомогою ДСЧО, розглядається функція приналежності, що враховує те, як проявляються на ДСЧО зміни частоти для різних людей і як відбуваються зміни в часі. Цей метод називають нечітким зіставленням образів [6].

За допомогою описаного вище методу розпізнавання була створена реальна система розпізнавання. Експерименти на цій системі проводилися японською, англійською та німецькою мовами. Японський набір включав 110 команд управління апаратурою для автоматизації установ, доповнений цифрами і звичайними словами [7], англійська та німецька – 120 слів такого ж змісту, а також назви тварин і квітів [8]. Результати розпізнавання, що отримані: японська мова – 93,2 %, англійська – 92,8 %, німецька – 95,7 %.

Відомі інші спроби використання нечіткої логіки в цих цілях. В працях [9, 10] представлені системи, побудовані на базі нечітких [9] та еволюційних нейро-нечітких систем [10]. Результати тестування цих систем показали схожу точність розпізнавання, як і в попередніх розглянутих працях.

Багато праць присвячено проблемі оцінки ваги потоку. Наприклад, в праці [11] оцінюються на базі алгоритму максимальної правдоподібності. В працях [12, 13] ваги потоку були розглянуті як параметри моделі і оцінені з використанням породжуючих або дискримінаційних критеріїв. У працях [14, 15] ваги потоку вважаються залежними і оцінюються для кожного кадру, ґрунтуючись на різних мірах надійності за допомогою евристично визначених функцій відображення, таких як сигмоїда або експоненціальна функція. Однак в цих працях не показано, чи є ці функції оптимальним вибором. У даному дослідженні пропонується аудіовізуальна система з неявним вибором відповідної функції зіставлення.

3. Мета та задачі дослідження

Мета дослідження – створення аудіовізуальної системи розпізнавання мови.

Для досягнення мети дослідження необхідно було вирішити наступні задачі:

1. Розробити загальну структуру аудіовізуальної системи розпізнавання мови.
2. Розробити класифікатори аудіо та відео компонент мови.
3. Дослідити побудовану систему на якість розпізнавання мови.

4. Аудіо-відео системи розпізнавання

Проаналізувавши алгоритми і моделі розпізнавання з розглянутих вище праць була запропонована наступна структура моделі всієї системи розпізнавання мови (рис. 1).

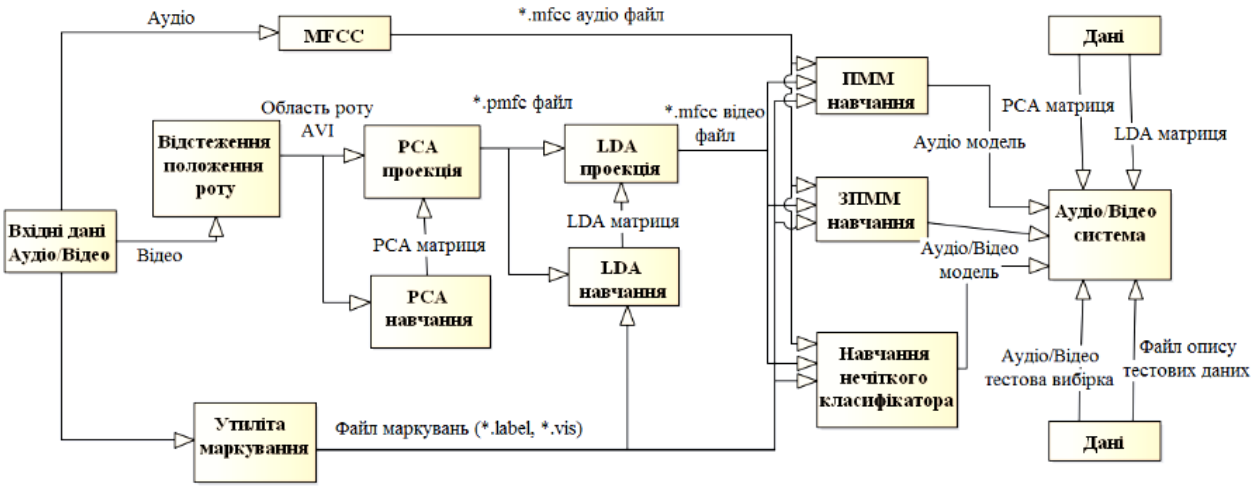

Рис. 1. Загальна структура моделі системи розпізнавання мови

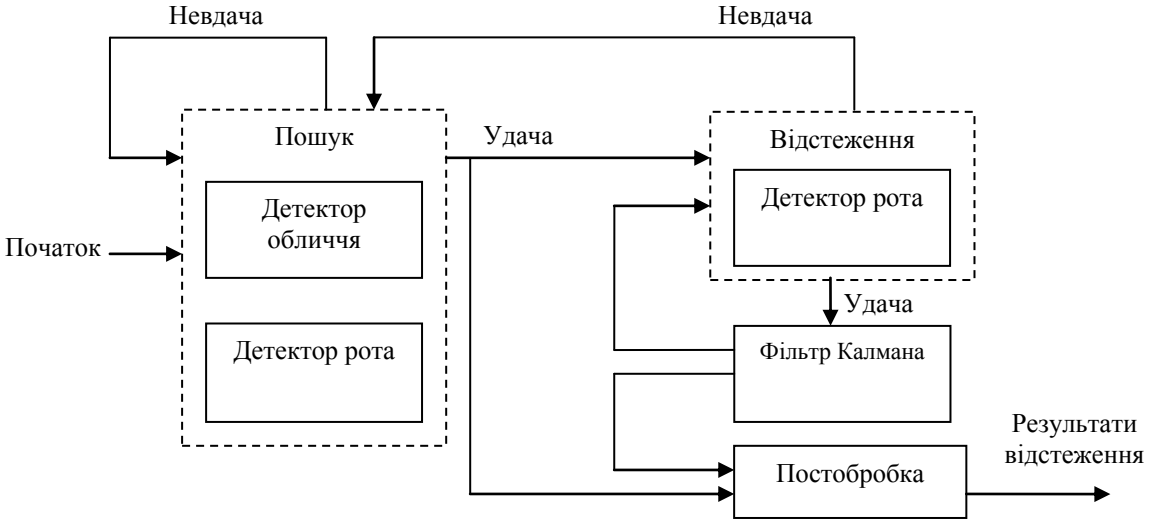

Рис. 2. Структура моделі для пошуку та відстеження області рота

Розглянемо основні елементи системи більш детально.Для пошуку і відстеження області рота була розроблена наступна модель (рис. 2).

Як видно з рис. 2, ядром моделі є автомат з двома станами: пошук і відстеження.

Система починає багатомасштабний пошук області обличчя за допомогою форсованого каскадного класифікатора, побудованого згідно [16] з використанням ознак Хаара. Далі, двокаскадний класифікатор (один каскад для області рота, інший – для області рота з підборіддям) відшукує область рота в нижній ділянці обличчя. Якщо обличчя знайдено в декількох послідовних кадрах, то автомат переходить в стан відстеження.

У режимі відстеження алгоритм детектування рота застосовується до маленької області навколо передбаченого положення області рота з попередніх кадрів.

Центр області пошуку вираховується за допомогою лінійного фільтра Калмана [17].

Оцінка положення рота на будь-якому часовому відрізку є плавною величиною, тому різкі скачки оцінки відкидаються за допомогою триетапної постобробки. На першому етапі виконується лінійна інтерполяція для заповнення прогалин в траєкторії руху області рота в зв'язку з помилками пошуку. На другому етапі застосовується медіанний фільтр для виключення неправильного детектування. На останньому етапі застосовується фільтр Гауса для придушення ефекту тремтіння траєкторії руху області рота.

Отримана за допомогою функції відстеження області рота послідовність зображень цих же облас-

тей нормалізується до розміру 32×32 пікселів і подається на каскад видобування ознак (рис. 3). Перш за все, зображення області рота відображається в 32-мірний простір ознак за допомогою функції аналізу головних компонент (PCA). Далі, для векторної послідовності підвищується дискретизація до 100 Гц з метою відповідності аудіальній ознаці і після цього, векторна послідовність проганяється через алгоритм нормалізації усереднених ознак [18]. Далі, всі вектори ознак об'єднуються в один вектор ознак за допомогою операції конкатенації. І нарешті, вектор ознак обробляється за допомогою методу аналіз головних компонент (LDA) на основі візем. На виході каскаду перетворень отримуємо вектор візуального спостереження.

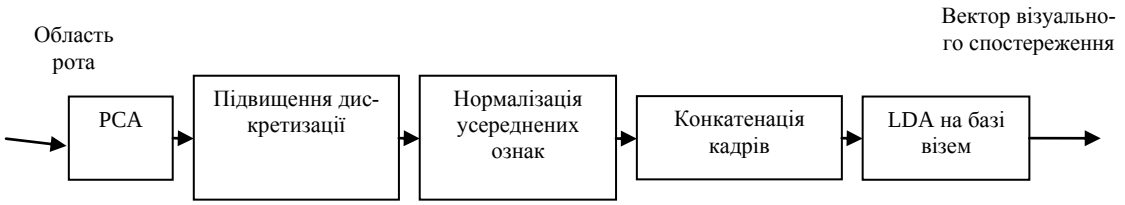

Рис. 3. Структура моделі видобування візуальних ознак

Модель класифікатора на базі ПММ. Акустичними одиницями моделювання є елементи вимови слів, які називаються фонемами (монофонія). Речення (набір слів) представляється як об'єднана послідовність фонем кожного слова. Для моделювання ефекту зчленування мови використовуються контекстні трифони.

На рис. 4 показана модель навчання, що складається з наступних 3 етапів:

1) фонемного навчання;

2) контекстного трифонного навчання;

3) кластерного трифонного навчання.

Розпізнавач мови побудований на базі алгоритму пошуку Вітербі (пошук найкращої послідовності станів, що відповідає даній вимові мови). Для роботи розпізнавача потрібні:

1) акустична модель для підбору акустичних даних;

2) мовна модель для визначення синтаксису і семантики;

3) словник вимов для правильної організації ПММ при пошуку.

Структура моделі пошуку показана на рис. 5. На виході розпізнавача мови знімається або транскрипція мови, або граф слів, або і те, і інше.

Модель класифікатора на базі ЗПММ. Зв'язані приховані Марківські моделі (ЗПММ) можуть розглядатися як набір звичайних ПММ, в якому кожна ПММ використовується для одного потоку даних і де приховані основні вузли часу t для кожної ПММ залежать від стану основних вузлів часу t–1 всіх ЗПММ. На рис. 6 представлена модель двопотокової ЗПММ для системи аудіовізуального розпізнавання мови.

Квадратиками на рисунку позначені приховані дискретні вузли (основні і змішані вузли), кружечками – вузли, що постійно спостерігаються. На відміну від ПММ, які використовуються для аудіо-та відеоданих, ЗПММ має можливість фіксувати взаємодію між аудіо- та відеопотоками за допомогою передачі ймовірностей між основними вузлами. ЗПММ має можливість моделювати аудіовізуальний стан асинхронно і таким чином, зберегти звичайну залежність між аудіо і відео в часі.

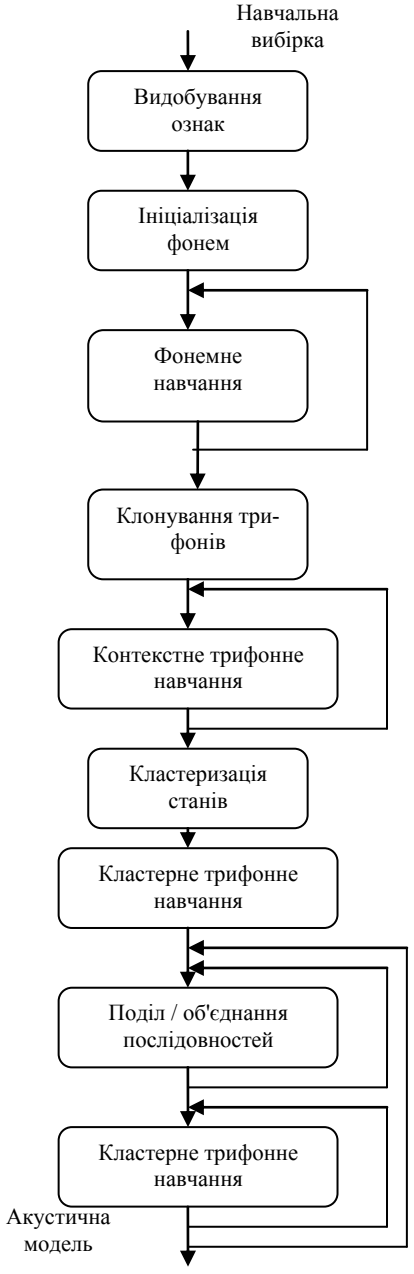

Рис. 4. Структура моделі навчання

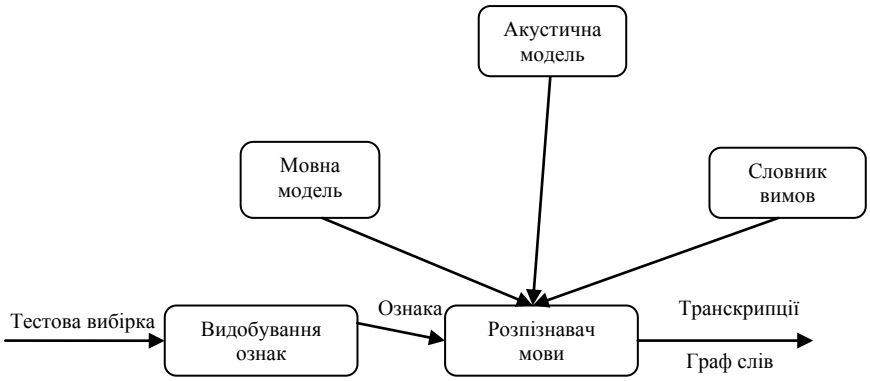

Рис. 5. Структура моделі тестування

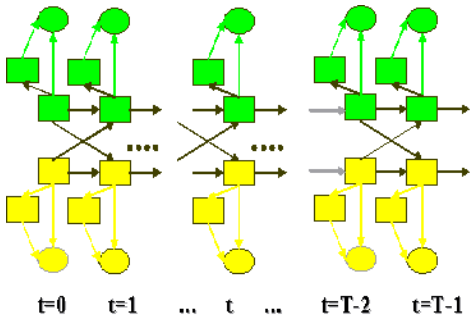

Рис. 6. Структура моделі двопотокового класифікатора на базі ЗПММ

Навчання ЗПММ параметрів виконується в два етапи. На першому етапі, параметри ЗПММ обчислюються для ізольованих пар фонема-візема. Параметри ізольованих пар фонема-візема ЗПММ спочатку оцінюються за допомогою алгоритму ініціалізації Вітербі [2] і далі проганяються через алгоритм оцінювання/максимізації (ЕМ) [19]. На другому етапі параметри ЗПММ, обчислені індивідуально на першому етапі, уточнюються за допомогою вкладеного навчання всіх ЗПММ. Схожим способом вкладеного навчання для ПММ [20], до кожної моделі, отриманої на першому етапі, додається один вхідний і один вихідний стани.

Аудіовізуальне розпізнавання виконується за допомогою декодера графів, який застосовується до мережі слів, що складається з усіх слів тестового словника. Кожне слово в мережі зберігається як послідовність фонем-візем ЗПММ, і найкраща послідовність слів визначається за допомогою розширеного алгоритму передачі маркера [20].

5. Експериментальні результати

Запропонована аудіовізуальна система розпізнавання була протестована на базі цифр від 0 до 9. Кожна цифра в базі даних повторюється десять разів кожним з десяти мовців. Для кожного оратора дев'ять прикладів кожної цифри використано для навчання, а приклад, що залишився, використано для тестування.

Усереднені результати розпізнавання для трьох режимів розпізнавання представлені в табл. 1. Режим розпізнавання наступні:

– тільки аудіо;
– тільки відео;
– аудіовізуальне розпізнавання.

Таблиця 1

Усереднені результати розпізнавання

Класифікатор	Результат розпізнавання
СММ (тільки аудіо)	92 %
СММ (тільки відео)	51 %
ССММ (аудіовізульна система)	98 %
МСММ [18]	55 %

Для режиму розпізнавання мови «лише аудіо» вектори акустичного спостереження (15 MFCC коефіцієнтів, які визначені з вікна 20 мс) моделюються з використанням ПММ. Для розпізнавання в режимі «аудіо-відео» використана ЗПММ з п'ятьма станами для зв'язаних вузлів як в аудіо-, так і в відеопотоках, без зворотних переходів і трьома змішуваннями на кожний стан.

Експериментальні результати показали, що рівень розпізнавання аудіовізуальної мови на основі ЗПММ збільшується на 43 % щодо розпізнавання мови в режимі «тільки аудіо». У порівнянні з багатопотоковою ПММ [18] запропонований варіант ЗПММ для аудіовізуальної системи розпізнавання показує кращі результати.

6. Висновки

1. Розроблено загальну структуру аудіовізуальної системи розпізнавання мови.

2. Розроблено класифікатори аудіо та відео компонент мови на базі прихованих Марківських моделей.

3. Досліджено побудовану систему на якість розпізнавання мови на прикладі аудіовізуального розпізнавання цифр.

Побудована система показала якість розпізнавання на рівні 98 %.

В подальшому планується розглянути алгоритми зважування потоків з метою покращення загальної розпізнавальної здатності.

Література

1. Liang, L. Speaker independent audio-visual continuous speech recognition [Text] / L. Liang, X. Liu, Y. Zhao, X. Pi, A. V. Nefian // International Conference on Acoustics, Speech and Signal Processing. – Lausanne, 2002. doi: 10.1109/icme.2002.1035365

2. Nefian, A. V. An coupled hidden Markov model for audio-visual speech recognition [Text] / A. V. Nefian, L. Liang, X. Pi, X. Liu, C. Mao // International Conference on Acoustics, Speech and Signal Processing. – Lausanne, 2002. doi: 10.1109/icassp.2002.1006167

3. Liang, L. Audio-Visual continuous speech recognition using a coupled hidden markov models [Text] / L. Liang, X. Liu, Y. Zhao, X. Pi, A. V. Nefian // International Conference on Acoustics, Speech and Signal Processing. – Lausanne, 2002. doi: 10.1109/icassp.2002.1006166

4. Gurban, M. Audio-visual speech recognition with a hybrid SVM-HMM system [Electronic resource] / M. Gurban, J. P. Thiran // 13th European Signal Processing Conference. – 2005. – Available at: https://infoscience.epfl.ch/record/87309/files/Gurban2005_1391.pdf

5. Raskinis, G. Building Medium-Vocabulary Isolated-Word Lithuanian HMM Speech Recognition System [Text] / G. Raskinis, D. Raskinien˙e // Informatica. – 2003. – Vol. 14, Issue 1. – P. 75–84.

6. Kass, M. Snakes: Active contour models [Text] / M. Kass, A. Witkin, D. Terzopoulos // International Journal of Computer Vision. – 1988. – Vol. 1, Issue 4. – P. 321–331. doi: 10.1007/bf00133570

7. Rao, R. R. Lip modeling for visual speech recognition [Text] / R. R. Rao, R. M. Mesereau // 28th Annual Asilomar Conference on Signals, Systems, and Computers. – 1994. – Vol. 1. – P. 587–590. doi: 10.1109/acssc.1994.471520

8. Sanchez, M. U. R. Statistical chromaticity-based lip tracking with B-splines [Text] / M. U. R.Sanchez, J. Matas, J. Kittler // IEEE International Conference on Acoustics, Speech and Signal Processing. – Munich, 1997. doi: 10.1109/icassp.1997.595416

9. Malcangi, M. Audio-visual fuzzy fusion for robust speech recognition [Text] / M. Malcangi, K. Ouazzane, P. Patel // The 2013 International Joint Conference on Neural Networks (IJCNN). – Dallas, 2013. doi: 10.1109/ijcnn.2013.6706789

10. Malcangi, M. Bio-inspired Audio-Visual Speech Recognition Towards the Zero Instruction Set Computing [Text] / M. Malcangi, H. Quan // International Conference on Engineering Applications of Neural Networks EANN 2016: Engineering Applications of Neural Networks. – 2016. – P. 326–334. doi: 10.1007/978-3-319-44188-7_25

11. Hernando, J. Maximum likelihood weighting of dynamic speech features for CDHMM speech recognition [Text] / J. Hernando // IEEE International Conference on Acoustics, Speech, and Signal Processing. – Munich, 1997. doi: 10.1109/icassp.1997.596176

12. Gravier, G. Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR [Text] / G. Gravier, S. Axelrod, G. Potamianos // IEEE International Conference on Acoustics Speech and Signal Processing. – Orlando, 2002. doi: 10.1109/icassp.2002.5743873

13. Peng, L. Stream weight training based on MCE for audio-visual LVCSR [Text] / L. Peng, W. Zuoying // Tsinghua Science and Technology. – 2005. – Vol. 10, Issue 2. – P. 141–144. doi: 10.1016/s1007-0214(05)70045-6

14. Estellers, V. On dynamic stream weighting for audio-visual speech recognition [Text] / V. Estellers, M. Gurban, J.-P. Thiran // IEEE Trans. Audio, Speech, and Language Processing. – 2012. – Vol. 20, Issue 4. – P. 1145–1157. doi: 10.1109/tasl.2011.2172427

15. Garg, A. Frame-dependent multi-stream reliability indicators for audio-visual speech recognition [Text] / A. Garg, G. Potamianos, C. Neti, T. S. Huang // International Conference on Multimedia and Expo. – Baltimore, 2003. doi: 10.1109/icme.2003.1221384

16. Lienhart, R. An extended set of Haar-like features for rapid objection detection [Text] / R. Lienhart, J. Maydt // Proceedings. International Conference on Image Processing. – Rochester, 2002. – P. 900–903. doi: 10.1109/icip.2002.1038171

17. Cordea, M. D. Real-time 2(1/2)-D head pose recovery for model-based video-coding [Text] / M. D. Cordea, E. M. Petriu, N. D. Georganos, D. C. Petriu, T. E. Whalen // IEEE Transactions on Instrumentation and Measurement. – 2001. – Vol. 50, Issue 4. – P. 1007–1013. doi: 10.1109/19.948316

18. Neti, C. Audio-visual speech recognition: Final Workshop 2000 Report, Center for Language and Speech Processing [Text] / C. Neti, G. Potamianos, J. Luettin et. al. – Baltimore: The Johns Hopkins University, 2000.

19. Jensen, F. V. An Introduction to Bayesian Networks [Text] / F. V. Jensen. – London: UCL Press Limited, 1998. – 178 p.

20. Young, S. The HTK Book [Text] / S. Young et. al. – Cambridge: Entropic Cambridge Research Laboratory, 1995.

Рекомендовано до публікації д-р техн. наук Телеником С. Ф.
Дата надходження рукопису 26.10.2017

Горносталь Олександр Миколайович, кафедра автоматики і управління в технічних системах, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», пр. Перемоги, 37, м. Київ, Україна, 03056
E-mail: gornostal.alexandr@gmail.com

Дорогий Ярослав Юрійович, кандидат технічних наук, доцент, кафедра автоматики і управління в технічних системах, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», пр. Перемоги, 37, м. Київ, Україна, 03056
E-mail: cisco.rna@gmail.com