Технічні науки

УДК 004.934.8

Малишев Андрій Ігорович
студент
Національний технічний університет України
«Київський політехнічний інститут ім. Ігоря Сікорського»
Малышев Андрей Игоревич
студент
Национальный технический университет Украины
«Киевский политехнический институт им. Игоря Сикорского»
Malyshev Andrii
student
National technical university of Ukraine
«Igor Sikorsky Kyiv Polytechnic Institute»

ВИБІР ЕФЕКТИВНОЇ ВІДКРИТОЇ СИСТЕМИ РОЗПІЗНАВАННЯ МОВЛЕННЯ В РЕЖИМІ РЕАЛЬНОГО ЧАСУ
ВЫБОР ЭФФЕКТИВНОЙ ОТКРЫТОЙ СИСТЕМЫ РАСПОЗНАВАНИЯ РЕЧИ В РЕЖИМЕ РЕАЛЬНОГО ВРЕМЕНИ
CHOISING OF EFFECTIVE OPEN SOURCE REAL-TIME SPEECH RECOGNITION SYSTEM

Анотація. В даній роботі описано структуру мовлення та принципи її розпізнавання. Описано відкриті системи розпізнавання та обрано оптимальну для розробки алгоритму синхронізації анімації обличчя віртуального персонажу із звуком.

Ключові слова: Розпізнавання голосу, віртуальна реальність, приховані моделі Маркова, фонеми, акустична модель, фонетичний словник.

Аннотация. В данной работе описана структура речи и принципы ее распознавания. Описаны открытые системы распознавания и избрана

оптимальная для разработки алгоритма синхронизации анимации лица виртуального персонажа со звуком.

Ключевые слова: Распознавание голоса, виртуальная реальность, скрытые модели Маркова, фонемы, акустическая модель, фонетический словарь.

Summary. This paper describes the structure and principles of speech recognition. Described and selected open recognition system for synchronization of facial animation and voice of the character in virtual reality.

Keywords: voice recognition, virtual reality, hidden Markov model, phoneme acoustic model, phonetic dictionary.

Вступ

Автоматичне розпізнавання мовлення являє собою актуальну задачу, зв'язану з безліччю різних сфер. Разом з тим, як стрімко набирає популярність таке поняття, як віртуальна реальність - розпізнавання мовлення людини стає невід'ємною частиною і актуальною проблемою її соціальної складової. При цьому якість звукового потоку може сильно відрізнятися в залежності від середовища, в якому знаходиться людина.

Крім того, людина яка розмовляє, може запинатися, змінювати темп мови. Так як ці фактори впливають на сприйняття мови людиною, звичайно ж, що вони будуть впливати і на автоматичне розпізнавання. Додатковою проблемою є те, що постає все більша потреба обробки мовлення та отримання із неї тексту або структурних одиниць, наприклад, фонем в режимі реального часу. У даній роботі буде розлянуто загальний процес розпізнавання голосу у відкритих системах, таких як CMU Sphinx та Kaldi, а також визначено, яка з них краще підходить для використання в концепції соціальної віртуальної реальності.

Структура мовлення

Мова являє собою безперервний аудіопотік, який перебуває в різних змінних станах. Серед цих станів виділяються найбільш схожі класи,

фонеми. При цьому властивості звукової хвилі, яка відповідає даній фонемі, залежать від багатьох чинників, таких як контекст фонеми, мовні особливості людини і т.д. Більш того, так як переходи між звуками більш інформативні, ніж стійкі відрізки, дослідники виділяють дифони (diphones) - ділянки аудіосигналу між двома послідовними фонемами.

Часто фонеми розглядаються в деякому контексті. Такі фонеми називають трифони (triphones). Тобто звучання конкретної фонеми може відрізнятися в залежності від оточуючих її фонем. На відміну від дифонів, їм відповідають ті ж відрізки звукової хвилі, що і фонемам.

Для зручності краще виявляти частини трифонів, а повністю. Наприклад, можна побудувати детектор для початку якогось трифона і використовувати його для інших з таким же початком. Тобто можна створити безліч таких детекторів для дуже коротких звуків. В CMU Sphinx такі детектори називають сенони (senones).

З фонем формуються підслова (subwords), тобто склади. Хоча при пришвидшені мовлення фонеми можуть змінюватися, склади залишаються незмінними. Різні комбінації підслів формують слова. Слова і нелінгвістичні звуки, наприклад, зітхання чи кашель, формують вислови - ділянки звукового потоку між паузами[1,2].

Розпізнавання

Типовий процес розпізнавання відбувається наступним чином. Звукова хвиля розбивається на частини по ділянках з тишею. Залежно від системи і використовуваних даних встановлюється певний мінімум гучності, наприклад, 10 Дб, і час утримання цього рівня, наприклад, 100 мс, або 10 фреймів. Також може бути взяте до уваги зниження рівня гучності на певну кількість Дб. Існують і інші методи поділу аудіосигналу на частини, на яких присутня мова. Далі можливі комбінації слів зіставляються з аудіорядом, з яких вибирається найкраща. Аудіосигнал розбивається на фрейми (зазвичай, по 10 мс), і для кожного фрейму обчислюється вектор ознак. В системі CMU Sphinx даний ветор складається з 39 значень. Вони

залежать від методу отримання ознак. Пошук найбільш ефективного способу обчислення цих значень все ще є предметом вивчення, проте в загальних випадках обчислюється похідна від спектра. Далі, отримані вектори розглядаються в рамках певних моделей - мовної, акустичної та фонетичного словника - для підбору комбінації слів ,яка є найкращою в даному випадку.

Для розпізнавання шаблонів у мовленні застосовуються приховані моделі Маркова. Тому в даному випадку найбільш підходящою комбінацією буде найбільш вірогідна. У кожен момент часу підтримуються поточні найкращі комбінації, які розширюються за рахунок пошуку комбінацій для наступного фрейму[2,3].

Моделі опису мови, які застосовуються при розпізнаванні

У розпізнаванні мови використовують комбінацію наступних моделей:

1. Акустична модель - містить інформацію про звукові параметри сенонів. Акустичні моделі бувають залежні або незалежні від контексту. Позначимо набір ознак як O, тоді для слова W ймовірність зустріти набір O - P(O | W)

2. Фонетичний словник - містить представлення слів у вигляді фонем. Представлення у вигляді словника не є обов'язковим - це може бути будь-яка функція - результат машинного навчання. Позначимо вимову слова W як Q(W)

3. Мовна модель - встановлює ліміт на кількість слів для пошуку. Вона вказує, яке слово може слідувати за даними і з якою ймовірністю, таким чином відсікаючи неможливі варіанти. Найбільш часто використовуються n-грамні мовні моделі. Позначимо ймовірність отримати слово W як P(W). Взявши X як вихідний аудіосигнал, набір слів як W, то задача розпізнавання мови формалізується наступним чином:

$$W^* = \arg\max_{W \in \mathbf{W}} P(W \mid X) = \arg\max_{W \in \mathbf{W}} P(O \mid Q(W))P(W)$$

Огляд існуючих відкритих систем систем

Існує кілька відомих відкритих систем розпізнавання мови, які надають потрібний функціонал для використання у концепції віртуальної реальності: CMU Sphinx та Kaldi. Перша реалізована на мові Java (хоча деякі бібліотеки написані на C), друга - на C++. Їх робота заснована на прихованих моделях Маркова. Основні відмінності у алгоритмах та мовленнєвих моделях, які використовуються.

CMU Sphinx

CMUSphinx - система розпізнавання мови, яка об'єднує ряд інструментів для задач різної складності. Ці інструменти розроблялися спеціально для слабких систем. Застосовується більше в прикладних цілях, а не в дослідженнях. На даний момент реалізована підтримка американської і британської англійської, французької, німецької, голландської і російської мов. Наявна потужна спільнота та підтримка користувачів.

Вибір системи CMU Sphinx для робіт зв'язаних із віртуальною реальністю є найкращим рішенням та обумовлений великим об'ємом робіт на її основі, а також суб'єктивним фактором - зручністю користування[4].

Kaldi

Основні напрямки використання і розвитку Kaldi - дослідження в області розпізнавання мови. Основна выдмінність від інших систем - інтеграція Finite State Transducers (FSTs) і широке використання лінійної алгебри. Підтримку FST забезпечує бібліотека з відкритим вихідним кодом OpenFST, лінійної алгебри - бібліотеки BLAS1 (Basic Linear Algebra Subroutines) і LAPACK2 (Linear Algebra PACKage). Як і CMU Sphinx, Kaldi надає вихідний код у відкритій формі, що дозволяє використовувати його для великого спектру завдань. Також може використовуватися для створення складних систем розпізнавання мови[5].

Висновки

Вибір системи розпізнавання мовлення був проведений для розробки алгоритму синхронізації анімації обличчя віртуального персонажа із звуком. Для даного завдання потрібна можливість розпізнавання у реальному часі, висока точність, методи отримання фонем та їх тривалостей із аудіопотоку.

У даній роботі була розглянута система CMUSphinx, яка має високі показники точності на даних високої якості. Оцінка відбувалась за показником WER(Word Error Rate). Чим вище WER, тим гірше відбулось розпізнавання. Зазвичай виражається у відсотках.

Показники WER(Word Error Rate) для не дуже зашумленого аудіопотоку виявились близькі до 30%, що є прийнятно для задач, які поставлені. Результати WER для зашумленого і дуже зашумленого аудіопотоку з різними особливостями вимови майже 50% - є дуже високими і не задовільняють вимогам. Отже, необхідно виробляти додаткову обробку аудіосигналу для зменшення рівня шуму. Також, при адаптації акустичної моделі, показники WER зменшувались[4,5].

Література

1. K.-F. Lee. Context-dependent phonetic hidden Markov models for speakerindependent continuous speech recognition.
2. George E. Dahl, Dong Yu, Li Deng, Alex Acero. Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition.
3. L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition.
4. CMU Sphinx Project by Carnegie Mellon University. http://cmusphinx.sourceforge.net/
5. Kaldi ASR. http://kaldi-asr.org/