Технічні науки

УДК 004.832

Хіхловська Наталія Олександрівна
бакалавр акустотехніки
Національного технічного університету України
"Київський політехнічний Інститут імені Ігоря Сікорського"
Хихловська Наталья Александровна
бакалавр акустотехники
Национального технического университета Украины
"Киевский политехнический институт имени Игоря Сикорского"
Khikhlovska Nataliia
Bachelor of Acoustic Equipment of the
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"

КЛЮЧОВІ ПРОБЛЕМИ СУЧАСНИХ СИСТЕМ ОЗВУЧУВАННЯ ТЕКСТОВОЇ ІНФОРМАЦІЇ

Анотація. *У цій статті визначено проблеми, які виникають при спробах досягти високої якості синтезу мовлення. Розглянуто чотири основні та дві додаткові проблеми. Останні в майбутньому будуть центральними для подальшого дослідження систем озвучування текстової інформації. Описано шляхи вирішення наступних задач: семіотична класифікація текстових фрагментів, врахування синтаксичних особливостей тексту, створення натурального та розбірливого звучання, створення емоційної забарвленості.*

Ключові слова: *системи озвучення текстової інформації, синтез мовлення, просодія, семіотичні системи, розбірливість.*

Аннотация. *В этой статье определены проблемы, которые возникают при попытках достичь высокого качества синтеза речи. Рассмотрены четыре основные и две дополнительные проблемы. Последние в будущем будут центральными для дальнейшего исследования систем озвучивания текстовой информации. Описаны пути решения следующих задач: семиотическая классификация текстовых фрагментов, учет синтаксических особенностей текста, создание натурального и разборчивого звучания, создание эмоциональной окрашенности.*

Ключевые слова*: системы озвучивания текстовой информации, синтез речи, просодия, семиотические системы, разборчивость.*

Summary. *This article describes the problems that arise when trying to achieve high-quality speech synthesis. Considered four main and two additional problems. The last one will be central for further research of text-to-speech systems. Described solutions to such problems: semiotic classification of text fragments, syntactic features of the text, creation of natural and distinctive voice, creation of prosody.*

Key words: *text-to-speech systems, speech synthesis, prosody, semiotic systems, legibility.*

Вступ. Для створення систем автоматичного озвучування текстової інформації перш за все потрібно визначити актуальні питання, та шляхи їх вирішення. Метою цієї статі є висвітлити основні проблеми створення систем синтезу мовлення. На основі аналізу літературних джерел визначено чотири основні напрями роботи. Визначено на якому рівні ці питання вже вирішено, та які ще дослідження необхідно зробити, щоб вдосконалити сучасні системи озвучування текстової інформації. Також освітлено дві додаткові проблеми,

які раніше не були вирішенні для таких систем.

1. Класифікація тексту відповідно до семіотичних систем

Помилково вважати, що текст завжди кодує природну мову. Скоріше, ми повинні бачити текст як загальний фізичний сигнал, який може бути використаний для кодування багатьох різних семіотичних систем, з яких природна мова є лише одним досить спеціальним випадком.

Існує два основні способи вирішення цієї проблеми. Перший підхід - нормалізація тексту, згідно з яким система бачить текст як інформацію на вході синтезатора і намагається переписати будь-який «нестандартний» текст як відповідний «лінгвістичний» текст. Другий - класифікувати кожну частину тексту відповідно до одного з відомих семіотичних класів. Звідси, алгоритм, специфічний для кожного класу, використовується для аналізу відповідної частини тексту і розкриття змісту. Для природної мови завдання аналізу тексту виконано, але для інших систем необхідний додатковий етап, де основна форма перекладається на слова.

Розглянемо семіотичний класовий підхід [1]. Ми можемо розділити речення на послідовність текстових лексем. Для кожної лексеми ми можемо визначити відповідний клас. Наприклад "21/09/2019" буде визначено як дата. Далі переводимо відповідну лексему у її змістовну форму. Наприклад "9:20am" ми можемо сформувати наступним чином: (години=9, хвилини=20, частина дня=ранок). Нарешті ми зможемо перевести це в слова. У мовленні є декілька різних прийнятних форм для вираження часу. У нашому прикладі можна сказати "двадцять хвилин на десяту ранку" або "дев'ята двадцять ранку", тощо. Може бути багато правильних варіантів, але оскільки система може читати тільки один, ми повинні бути чутливими до того, який з них має бути. Іноді перевага залежать від користувача. Крім того, контекст і жанр тексту можуть впливати на вибір варіанту. Ще однією проблемою є те, що для

деяких лексем немає узгодженого читання. Наприклад електронні адреси, комп'ютерні програми або інша технічна інформація. Візьмемо електронну адресу "hi30@gmail.com". Виникає питання як слід її читати. Має бути "hi30" прочитано як "hi тридцять" або як "H I тридцять" або як "H I три нуль"?

Таким чином відносно даної проблеми можемо поставити три завдання:

1) Визначити до якого семіотичного класу відноситься лексема.
2) Проаналізувати кожну лексему та знайти її змістовну форму.
3) Перевести лексему у форму у якій вона буде прочитана.

2. Декодування тексту натуральної мови

Як тільки текст належним чином класифікований у термінах його семіотичного класу, загальний рівень неоднозначності щодо змісту значно зменшується. Але в області натуральної мови, досі залишається певна неоднозначність. Вона зустрічається в багатьох різних формах. Найбільшу проблему складають омоніми — слова, що однаково пишуться, але мають різне значення [2]. Також досить поширена синтаксична неоднозначність. Відомий приклад "стратити не можна помилувати", де значення можна трактувати двома різними способами в залежності від синтаксису: "стратити, не можна помилувати" або "стратити не можна, помилувати".

Питання при створенні системи озвучування текстової інформації - це те, наскільки багато цієї невизначеності нам потрібно вирішити. При комплексному підході необхідно було би вирішити всю неоднозначність у тексті і створити складну структуру для визначення змісту. Однак, чим більше ми намагаємося вирішити неоднозначності, тим частіше будемо робити помилку. Зважаючи на основний інженерний принцип ефективності [3], слід обрати мінімалістичний підхід, згідно з яким, ми повинні вирішувати тільки ті двозначності, які впливають на вимову. Наприклад слово "курс" використовується в українській мові багатьма різними способами, але воно

завжди написане однаково і завжди вимовляється однаково, тому ми можемо ігнорувати безліч значень і використань. Якщо ми знаємо, що певний текст може мати два або більше різних значення, але розуміємо, що всі ці форми вимовляються однаково, то ми не повинні намагатися вирішувати неоднозначність змісту.

3. Розбірливість

Третім основним завданням є розбірливість. Для наших цілей ми визначаємо розбірливість як здатність слухача декодувати повідомлення.

З усіх проблем систем синтезу мовлення розбірливість найлегше вирішити. Фактично, можна стверджувати, що цю проблему було вирішено досить давно, оскільки генерувати досить зрозумілі звучання мови в системах можна було з кінця 1970-х років. Насправді, в тестах може бути показано, що зрозумілість сучасних систем озвучування тексту часто є лише незначно мірою краще, ніж у набагато старших системах, таких як MITalk [6]. Але це не означає, що в цьому питанні не було досягнуто прогресу. Грубо кажучи, історія полягає в тому, що в той час як зрозумілість старих систем була достатньо високою, дослідження зосереджувалися на підвищенні природності. Тому дослідження полягало в тому, щоб поліпшити природність, не зробивши гіршою розбірливість.

Існує кореляція між природністю та розбірливістю. Для досягнення високої розбірливості більш ранні системи використовували дуже "безпечну" мову — тобто ігнорували плавність вимови, синтаксичні паузи, коартикуляцію звуків, тощо, що є дуже важливими аспектами у природності звучання. Оскільки ці ефекти додаються, існує ймовірність того, що вони призведуть до втрати зрозумілості. Тому хитрість полягає в тому, щоб додати додаткову природність, синтезуючи варіації більш доречно, але зробити це таким чином, щоб не призвести до втрати розбірливості.

4. Природність звучання

Головною метою дослідження систем синтезу мовлення є розробка системи, що буде давати на виході більш природне звучання. Під природним ми розуміємо, що система повинна звучати так само, як людина.

Чому важливо досягти природності звучання? Можна однозначно сказати, що більшість слухачів дуже нетерпимі до неприродності звучання, незалежно від мети використання [4]. В багатьох ситуаціях ми цілком задоволені візуальними карикатурами людей, про що свідчать карикатури та анімація. Наприклад ми не вважаємо малюнок Гомера Сімпсона жалюгідним перекладом реальності, але для нас є вкажливим, щоб він мав природний голос, тому в Сімпсонах (як і у всіх інших мультфільмах) для його голосу використовується справжній актор.

Що ж робить один голос більш природним, ніж інший? Знову ж таки, це питання, на яке сучасна наука не в змозі повною мірою відповісти, але ми можемо назвати декілька факторів.

По-перше, будь-яку систему, яка виражає очевидні нелюдські артефакти в мові, слід вважати неприродною. Тобто мовлення з клацаннями, гудінням і нескінченним набором інших механічних звуків.

По-друге, мовлення не повинно бути монотонним. Система має враховувати синтаксис речення, певне емоційне забарвлення. Тому, для забезпечення природності звучання, ми повинні зробити так, щоб синтезатор звучав як хтось: або копія голосу реальної людини, або штучно створений голос, який буде без проблем сприйнятий реальною людиною.

Для оцінки природності звучання наразі використовують MOS (Mean Opinion Score) [5]. Фактично результуючою оцінкою системи буде середнє значення оцінок фокус групи відносно суб'єктивного сприйняття натуральності звучання. При оцінюванні аудіо фрагменту згенерованого

системою виставляють оцінку у балах від 1 до 5, де 1 - дуже погано, 5 — відмінно. Сучасні системи озвучування текстової інформації за цією шкалою отримують оцінки близько 4.5 [8], тобто можна вважати що на сучасному етапі розвитку систем, досягнуті досить високі показники натуральності.

5. Допоміжні засоби для просодії

Тепер розглянемо першу з більш складних проблем - як створити алгоритм, що кодує весь спектр просодичних ефектів, а не нейтральну вимову характерну для більшості сучасних систем. Просодія не закодована в тексті, і тому ми ніколи не зможемо відновити її з написання. Як же нам генерувати просодичний зміст, якщо ми не можемо розкрити його з тексту?

По-перше, відокремимо питання генерування сигналу, який містить певний просодичний ефект (наприклад, несподіванку або акцентуацію конкретного слова) від питання про те, як зробити це автоматично з тексту. Для цього можна зібрати данні з відповідним ефектом і вивчити як сегментні ознаки змінюються залежно від нього.

Те, як автоматично визначати, які просодичні ефекти використовувати з тексту є більш складною проблемою. Це питання генерації, а не декодування тексту, і це робиться як окреме, допоміжне завдання для процесу вербального декодування і кодування. Неможливо однозначно визначити якою має бути просодія. Але, очевидно, що система, яка читає повідомлення, «ми з жалем повідомляємо вам про сумну смерть пана ...» з жартівливим або легковажним тоном не є досконалою. Майже завжди існує якийсь контекст, відповідно до якого ми можемо стверджувати, що певний акцент або настрій є більш адекватним.

Таким чином система має визначити імовірність відповідності певної просодії до конкретної частини тексту та застосувати найбільш імовірну для даного випадку.

6. Адаптація системи до ситуації

Питання тут полягає в тому, що на відміну від "нормальних" комунікативних ситуацій - читання мовчки або бесіди, у випадку читання вголос ми маємо три агенти: автор, слухач і читач (тобто система озвучування тексту). Більшість текстів ніколи не були написані з наміром читати їх вголос і через це можуть виникнути проблеми у тому, що послідовне читання тексту може призвести до ситуацій, коли читач говорить те, що слухач не може зрозуміти, не знає або не зацікавлений у слуханні. Коли люди беруть на себе роль читача, вони часто відхиляються від буквального тексту, використовуючи пояснення, перефразовують, використовують синоніми, тощо, щоб зрозуміти авторське послання. Наскільки адекватний читач до тексту, дуже сильно залежить від ситуації. Ми іноді описуємо хорошого читача, як людину, яка добре усвідомлює як слухача, так і наміри автора і зважаючи на це, може обирати компроміс між вірністю і прийнятністю [7].

Не часто цю проблему намагаються вирішити безпосередньо в системі озвучення тексту. Це зазвичай розглядається як окрема проблема. Незважаючи на це, систему синтезу мовлення можна вдосконалити додатковими функціями, такими як: зміна швидкості мовлення, повторення фраз, пропуск фраз, тощо.

Література

1. Д. Ділі, Основи семіотики ; пер. з англ. та наук. ред. А. Карася ; Львів. нац. ун-т ім. І. Франка, Філософ. ф-т. – 2-ге доп. вид. – Л. : Арсенал, 2000. – 232 с. – Тит. арк. парал. укр., англ. – Бібліогр.: С. 201-222. – ISBN 966-7790-00-2.
2. Шипнівська Ольга Визначення типів синтаксичної неоднозначності у знаннєорієнтованій системі машинного перекладу // Українське

мовознавство. — 2013. — № 43. — С. 104—113.

3. P. Taylor, Text-to-Speech Synthesis, Cambridge University Press, New York, NY, USA, 1st edition, 2009.
4. Helmholtz, Hermann von (1885), On the sensations of tone as a physiological basis for the theory of music, Second English Edition, translated by Alexander J. Ellis. London: Longmans, Green, and Co., p. 44. Retrieved 2010-10-12.
5. Streijl, Robert C., Stefan Winkler, and David S. Hands. "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives." Multimedia Systems 22.2 (2016): 213-227.
6. Allen J., Hunnicut S., and Klatt D., From Text to Speech: the MITalk System. Cambridge University Press, 1987.
7. Alwan A., Narayanan S., and Haker K., Towards articulatory-acoustic models for liquid consonants based on MRI and EPG data. part II: The rhotics. Journal of the Acoustical Society of America 101, 2 (1997), 1078–1089.
8. Jonathan Shen, Ruoming Pang, Ron J. Weiss, Natural Tts Synthesis By Conditioning Wavenet On Mel Spectrogram Predictions, Google, Inc., 2University of California, Berkeley, 2018.