УДК 004.93 DOI: 10.31673/2786-8362.2024.022953

Іщеряков С.М., к.т.н; **Попов А.О.**

ОЦІНКА ЯКОСТІ СИНТЕЗОВАНОГО МОВЛЕННЯ

Ishcheriakov S.M., Popov A.O. Evaluating synthesized speech quality. The article explores methods for objectively evaluating the quality of synthesized speech that can be used to assess the accuracy and performance of speech generation systems. This evaluation approach aims to provide an unbiased measure of the improvement of the quality of synthesized speech, independent of the subjective opinions of listeners, thus facilitating the design and improvement of speech synthesis systems. It also offers a framework for comparing different speech synthesis systems to determine which one performs better. Considerable attention is paid in the article to the use of neural networks as a tool for evaluating the quality of the output of other neural networks, emphasizing the potential of self-evaluation in artificial intelligence systems. The effectiveness of the existing quality control systems is thoroughly investigated with the determination of their strengths and weaknesses. In addition, the paper highlights the key advantages of each evaluation system available, contributing valuable information to the continuous improvement of speech synthesis technologies.

Keywords: neural network, synthesized speech, evaluation metrics, metrics.

Іщеряков С.М., Попов А.О. Оцінка якості синтезованого мовлення. У статті досліджується методи об’єктивної оцінки якості синтезованого мовлення, які можна використовувати для оцінки точності та продуктивності систем синтезу мовлення. Цей підхід до оцінювання має на меті забезпечити неупереджену міру підвищення якості синтезованого мовлення, незалежну від суб’єктивних думок слухачів, таким чином полегшуючи розробку та вдосконалення систем синтезу мовлення. Він також пропонує основу для порівняння різних систем синтезу мовлення, щоб визначити, яка працює краще. Значна увага в статті приділяється використанню нейронних мереж як інструменту для оцінки якості вихідного результату інших нейронних мереж, підкреслюючи потенціал самооцінки в системах штучного інтелекту. Ефективність існуючих систем контролю якості ретельно досліджується з визначенням їх сильних і слабких сторін. Крім того, стаття висвітлює ключові переваги кожної доступної системи оцінювання, вносячи цінну інформацію в постійне вдосконалення технологій синтезу мовлення.

Ключові слова: нейронна мережа, синтез мовлення, метрики оцінки, метрики.

Вступ

Синтез мовлення – це технологія, що постійно розвивається, еволюціонуючи від простих математичних систем до сучасних загорткових нейронних мереж. Якщо порівняти сьогоднішні системи синтезу мовлення з тими, які існували двадцять років тому, прогрес очевидний. Сучасні розробки важко порівнювати між собою, оскільки вони звучать майже ідентично людському мовленню, тому оцінка якості цих систем на слух стає все більш складною задачею. Прийняття рішення про те, яка система працює краще, лише через суб'єктивне сприйняття звукових доріжок не дає об'єктивних результатів і значно ускладнює процес оцінки. У відповідь на ці виклики пропоную проаналізувати існуючі методи об’єктивної та головне, автоматизованої оцінки якості мовлення, що дозволять стандартизувати підходи до порівняння різних систем.

Щоб створити систему оцінки, потрібно чітко визначити критерії, за якими ми будемо оцінювати результати синтезу мовлення. Аналіз наявних методів дозволить виділити найефективніші з них, що здатні забезпечити об’єктивність і точність. Головна ідея полягає в тому, щоб вибрати такі метрики, які не лише відповідатимуть реальній якості мовлення, а й будуть легко інтегровані в автоматизовані системи. Це дозволить усунути суб'єктивність людського сприйняття і створити стандарти, що забезпечать високу якість мовлення в майбутньому.

Постановка завдання. Науковим завданням є визначення об'єктивних метрик для оцінки якості синтезованого мовлення, які можуть бути використана для подальшого вдосконалення систем тексту в мовлення. Тому необхідно провести аналіз методів, що дозволять автоматично і точно оцінювати ключові аспекти синтезованого мовлення, такі як фонетична коректність,

розбірливість, інтонаційна і просодична природність, а також відповідність текстовій інформації.

Для цього потрібно:

– проаналізувати існуючі інструменти та підходи для оцінки якості мовлення;

– визначити та систематизувати метрики, що оцінюють якість мовлення з різних аспектів: акустичних, лінгвістичних, перцептивних та нейромережевих;

– зробити огляд сучасних моделей на основі глибокого навчання для автоматизованої оцінки якості синтезованого мовлення, таких як MOSNet, STOI-Net, Wav2Vec та BERTScore;

– запропонувати шляхи оптимізації процесу оцінки мовлення для його подальшого використання в системах синтезу мовлення.

Таким чином, необхідно створити надійну й універсальну систему метрик, яка дозволить автоматично оцінювати якість мовлення з урахуванням усіх ключових аспектів, потрібних для якісної синтезу мовлення.

Аналіз останніх досліджень. Аналіз останніх наукових публікацій показує, що науковці та розробники з всього світу активно досліджують сферу об’єктивної оцінки та покращення якості синтезу мовлення. Провідним теологічними рішеннями даної проблеми є використання нейронних мереж.

Так наприклад у статті [11] згадується необхідність якісної оцінки мовлення в процесі покращення нейронних мереж синтезу мовлення. В статті [12] обговорюються плюси та мінуси існуючих метрик оцінки якості, таких як NISQA та DNSMOS. В публікації [13] згадується корисність метрик у сфері виявлення синтезованого мовлення в безпекових цілях. Стаття [14] розкриває тему створення нейронних мереж для оцінки якості синтезованого мовлення навчених на оцінках реальних людей людей. У статті [15] приводиться порівняльна таблиця в який розглянуті основні якісні характеристики популярних нейронних мереж синтезу мовлення, які будуть згадані у даній статті.

На основі аналізу наукових публікацій та літературних даних було зроблено висновок, що перспективною технологією для вдосконалення сучасних нейронних мереж синтезу є створення метрик на основі нейронних мереж у поєднанні з існуючими метриками інших типів. Оскільки існуючі метрики є вузько направленими і тому охоплюють малу частину аспектів якості цікавих для вдосконалення систем синтезу. А сам процес вдосконалення існуючих метрик є необхідною та перспективною науковою задачею.

Метою роботи є підвищення ефективності оцінки якості синтезованого мовлення. В статті будуть проаналізовані інструменти, програми та техніки які можна використати для оцінки і вдосконалення системи синтезу мовлення.

Виклад основного матеріалу дослідження.

До основних метрик оцінки якості синтезованого мовлення відносять наступні:

– акустичні;
– лінгвістичні та просодичні;
– метрики перцептивних моделей;
– непромережених метрики;
– метрики узгодження контексту та мовної семантики;
– емоційні та виразні метрики;
– комбіновані метрики.

Акустичні метрики базуються на аналізі фізичних характеристик аудіо сигналу та порівнянні синтезованого мовлення з реальним зразком:

– PESQ (Perceptual Evaluation of Speech Quality): використовується для оцінки якості зв'язку й мовлення. Порівнює синтезоване мовлення з оригінальним на основі психоакустичних характеристик, таких як інтенсивність і частотні спектри;

– STOI (Short-Time Objective Intelligibility): метрика, яка вимірює розбірливість синтезованого мовлення. Вона оцінює наскільки легко слухачу зрозуміти синтезоване мовлення, особливо в умовах шуму;

– Mel-Cepstral Distortion (MCD): вимірює різницю між спектральними характеристиками синтезованого мовлення і реального запису. Мета — оцінити, наскільки відтворення мелодійної складової мовлення в синтезі відповідає оригіналу;

– Spectrogram Correlation: спектрограми синтезованого мовлення порівнюються з оригінальними з точки зору схожості між ними.

Хоча ці метрики оцінюють технічні аспекти мовлення, проте вони не завжди точно відображають сприйняття мовлення людиною. Для синтезу мовлення, яке буде використовуватись людьми, необхідно враховувати сприйняття слухачів, а не лише технічні характеристики сигналу.

Оцінка на основі лінгвістичних і просодичних характеристик:

– Intonation and Pitch Accuracy: об'єктивні метрики можуть оцінювати відповідність інтонації синтезованого мовлення до натурального. Наприклад, використання висоти тону (pitch) і частотної варіації (intonation) для порівняння синтезованого голосу з еталонним;

– Durational Metrics: вимірювання тривалості фонем, слів і пауз в синтезованому мовленні. Тривалість звуків важлива для передачі природності й розбірливості;

– Rhythm Matching: визначення, наскільки ритм синтезованого мовлення відповідає очікуваному для конкретної мови.

Перцептивні моделі намагаються відтворити те, як людський слух сприймає мовлення, і використовуються для автоматизованої оцінки якості:

– PLDA (Perceptual Linear Predictive Distortion): вимірює спотворення мовлення на основі того, як слухачі сприймають якість мовлення через математичні моделі сприйняття слуху;

– PER (Phone Error Rate): вимірює відсоток фонетичних помилок у синтезованому мовленні відносно правильного мовлення. Це дозволяє оцінити, наскільки розбірливо й чітко були синтезовані окремі фонеми.

Ці метрики більш орієнтовані на сприйняття людиною і можуть краще відображати суб'єктивні аспекти якості мовлення, зокрема, натуральність та емоційність.

Метрики на основі нейромереж. Нейромережі можна навчати для оцінки якості мовлення, ґрунтуючись на великих наборах даних реальних оцінок якості від людей. MOSNet (Mean Opinion Score Network): мережа, яка прогнозує середню оцінку якості мовлення (MOS) на основі аудіоданих. Навчена на суб'єктивних оцінках людей, MOSNet може передбачати людську реакцію на синтезоване мовлення. Wav2Vec: використовується для аналізу синтезованого мовлення й порівняння його з природними зразками за допомогою векторного подання звуку. Цей метод дозволяє оцінити якість мовлення з урахуванням як фонетичних, так і просодичних аспектів.

Метрики узгодження контексту та мовної семантики: Для моделей TTS важливо не лише синтезувати мовлення з правильною інтонацією і чіткістю, але й забезпечити адекватне синтаксичне і семантичне відтворення тексту.

– BLEU Score: Використовується для оцінки якості текстової частини синтезованого мовлення. Оцінює схожість між синтезованими та реальними текстовими транслітераціями;

– BERTScore: Визначає семантичну відповідність між синтезованим мовленням і оригінальним текстом за допомогою трансформерів на основі архітектури BERT. Це дозволяє оцінити, наскільки правильно синтезоване мовлення передає зміст тексту.

Емоційні та виразні метрики:

– Emotion Recognition Systems: можуть використовуватися для визначення, наскільки точно синтезоване мовлення передає емоції, такі як радість, сум, гнів тощо. Це може бути корисно для оцінки якості мовлення в емоційно насичених діалогах;

– Prosody Matching: просодія є важливим аспектом виразного мовлення. Метрики можуть вимірювати, наскільки добре синтезоване мовлення відповідає бажаним просодичним характеристикам, зокрема, змінюється тон і темп відповідно до контексту.

Комбіновані метрики:

– об'єднання різних об'єктивних метрик може дати більш точну оцінку якості мовлення. Для цього можна створювати гибридні моделі, які враховують як акустичні, так і семантичні аспекти мовлення;

– моделі на основі машинного навчання можна навчати, використовуючи різні метрики і людські оцінки якості, щоб отримати комплексний показник, який відповідає перцепції.

Для оцінки якості мовлення слід зосередити увагу на його основних аспектах: фонетичному, граматичному, лексичному, семантичному, комунікативному, прагматичному та інтонаційному. Важливо, щоб мовлення було чітким, правильно структурованим і відповідало комунікативній ситуації. Технічні аспекти мовлення можна перевіряти за допомогою акустичних метрик, таких як PESQ, STOI, MCD та спектрограмна кореляція. Однак, оскільки вони не завжди відображають сприйняття людиною, варто використовувати також лінгвістичні та просодичні оцінки, наприклад, точність інтонації, ритму та тривалості звуків.Перцептивні моделі, такі як PLDA та PER, краще відображають суб'єктивне сприйняття мовлення, включно з його натуральністю та емоційністю. Використання нейромережевих моделей, таких як MOSNet і Wav2Vec, допомагає оцінити як якість мовлення загалом, так і його фонетичні та просодичні характеристики. Крім цього, для моделей синтезу мовлення важливими є метрики, що оцінюють контекст та мовну семантику, наприклад, BLEU Score і BERTScore. Емоційні та виразні метрики можуть вимірювати передачу емоцій та відповідність просодичних характеристик. Комбіновані метрики та моделі, що враховують як технічні, так і перцептивні аспекти, дозволяють створити найбільш повну оцінку якості мовлення.

Таким чином, вимірювання якості синтезованого мовлення за допомогою нейронних мереж є найкращим, оскільки дозволяє охопити найбільшу кількість аспектів мовлення, тому розглянемо такі нейронні мережі більш детально.(Mean Opinion Score Network) — одна з перших спроб автоматизувати оцінку якості мовлення на основі людських оцінок, вимірюваних за шкалою MOS (Mean Opinion Score). MOS традиційно використовується для оцінки якості мовлення слухачами, але цей процес дорогий і тривалий. MOSNet розроблено для автоматичного прогнозування цих оцінок. Архітектура MOSNet передбачає вхід мовленнєвого сигналу (синтезованого аудіо), з якого витягуються спектральні особливості, такі як мел-спектрограми, для подальшої обробки нейронною мережею. Використовується стандартна глибока нейронна мережа або модель LSTM (Long Short-Term Memory), яка аналізує ці дані. Результатом є оцінка, що відповідає середньому значенню, яке могли б надати люди. MOSNet автоматизує процес оцінки мовлення, забезпечуючи швидкість оцінки без потреби залучення слухачів, що дозволяє оцінювати великі набори аудіофайлів. Проте оцінки не завжди відповідають суб’єктивному сприйняттю людей, оскільки модель навчена на обмеженому наборі даних. MOSNet здебільшого фокусується на натуральності мовлення, але не завжди здатна оцінити емоційність. STOI-Net (Short-Time Objective Intelligibility Network) — це нейромережевий підхід для оцінки розбірливості мовлення. STOI-Net не оцінює загальну якість, як це робить MOSNet, а визначає, наскільки чітко можна зрозуміти синтезоване мовлення. Вхідними даними для STOI-Net є аудіосигнал, з якого витягуються спектральні особливості, наприклад, мел-спектрограми. Використовуються моделі CNN або LSTM для обробки цих часових рядів. Результатом є оцінки, які відображають здатність людей зрозуміти мовлення. STOI-Net корисна для систем, де важлива розбірливість мовлення, таких як голосові помічники. Модель демонструє високу продуктивність у шумних середовищах, але вона фокусується лише на розбірливості, не враховуючи такі аспекти, як емоційність або інтонація.

DeepSpeech-QA — це модель для оцінки якості мовлення, заснована на принципах автоматичного розпізнавання мовлення (ASR). Вона не лише розпізнає мовлення, а й оцінює, наскільки точно синтезоване мовлення передає зміст тексту. На вході модель отримує синтезоване аудіо та відповідний текст, після чого аналізує їхню відповідність за допомогою нейронної мережі для розпізнавання мовлення. Оцінка якості вимірюється на основі точності розпізнавання та відповідності вхідному тексту. DeepSpeech-QA гарантує правильну передачу змісту тексту та може використовуватись для оцінки мовлення в реальному часі. Проте ця

модель зосереджена лише на точності передачі тексту, не враховуючи інші аспекти, як-от натуральність мовлення.

Wav2Vec – це потужна модель для обробки аудіосигналів, що використовує самонавчання. Вона витягує векторні подання аудіосигналів і може бути навчена для передбачення якості мовлення на основі цих подань. Ця модель дозволяє оцінювати різні аспекти мовлення, включно з розбірливістю, інтонацією та мелодійністю. Завдяки можливості працювати на необроблених даних, Wav2Vec забезпечує більш гнучкий підхід до оцінки мовлення, що може бути адаптований для різних мов і акцентів. Водночас модель потребує великих обсягів даних для навчання, а її оцінки можуть бути залежними від конкретного завдання або мовного середовища.

BERTScore зазвичай використовується для текстових завдань, але може бути адаптований для оцінки якості мовлення через порівняння семантичного змісту синтезованого мовлення з текстом, з якого воно було згенеровано. Вхідними даними є текст і синтезоване мовлення, а нейронна мережа BERT використовується для порівняння семантичних подань цих даних. Оцінка базується на семантичній відповідності тексту та мовлення.

BERTScore оцінює не лише правильність синтезу мовлення, але й його відповідність семантичному змісту, що робить його корисним для багатомовних систем або систем, де важливе точне передання змісту. Проте оцінка семантики не завжди корелює з натуральністю або емоційністю мовлення.

Розглянувши нейронні мережі для перевірки якості синтезованого мовлення можна зробити висновки, що оцінка якості синтезованого мовлення може здійснюватися за допомогою різних моделей та підходів, кожен з яких акцентує увагу на окремих аспектах мовлення. MOSNet автоматизує процес оцінки натуральності мовлення, швидко прогнозуючи оцінки на основі людських даних, але має обмеження в оцінці емоційності та виразності. STOI-Net, зосереджена на розбірливості мовлення, є ефективною в умовах шуму, проте не враховує інтонаційні та емоційні аспекти. DeepSpeech-QA забезпечує відповідність синтезованого мовлення вхідному тексту, що є важливим для гарантування точності передання змісту, однак не оцінює натуральність мовлення. Wav2Vec є гнучким інструментом для оцінки різних параметрів мовлення, включаючи інтонацію та мелодійність, але потребує значних обсягів даних для ефективного навчання. BERTScore забезпечує семантичну відповідність між текстом і мовленням, що робить його корисним для систем, де важливе точне передання змісту, хоча це не завжди корелює з натуральністю мовлення. Тому, комплексне оцінювання якості мовлення потребує використання кількох методів та моделей для врахування всіх важливих аспектів: розбірливості, натуральності, інтонації, емоційності та семантичної відповідності.

Висновки

Оцінка синтезованого мовлення повинна базуватися на ключових аспектах: фонетичному, граматичному, лексичному, семантичному, комунікативному, прагматичному та інтонаційному. Важливо, щоб мовлення було чітким, правильно структурованим та відповідало комунікативній ситуації. Для технічного аналізу якості мовлення використовуються акустичні метрики, такі як PESQ, STOI, MCD та спектрограмна кореляція. Однак, оскільки ці метрики не завжди відображають суб'єктивне сприйняття людиною, важливо також оцінювати мовлення з точки зору лінгвістичних і просодичних характеристик, таких як точність інтонації, ритм і тривалість звуків.

Кожен підхід до оцінки мовлення має свої переваги та обмеження. MOSNet добре підходить для оцінки натуральності мовлення, але може не враховувати емоційність, тоді як STOI-Net фокусується на розбірливості, що є важливим у шумних середовищах, проте не враховує інтонаційні аспекти. DeepSpeech-QA гарантує точне передання змісту тексту через синтезоване мовлення, але не забезпечує оцінку його натуральності. Wav2Vec є гнучкою моделлю, яка дозволяє оцінювати різні параметри мовлення, зокрема інтонацію та

мелодійність, але вимагає великих обсягів даних. BERTScore забезпечує семантичну відповідність між текстом і мовленням, що робить його корисним у багатомовних системах.

Для всебічної оцінки якості мовлення потрібно використовувати кілька методів, що поєднують технічні та суб'єктивні аспекти. Це дозволяє охопити всі важливі параметри: розбірливість, природність, інтонацію, емоційність і відповідність тексту, забезпечуючи цілісне розуміння якості синтезованого мовлення з урахуванням технічних характеристик сигналу та людського сприйняття.

Тому, для підвищення якості оцінки потрібно використовувати наступні метрики. Перцептивні моделі PLDA та PER, дозволяють краще відобразити суб'єктивне сприйняття мовлення, включно з його натуральністю та емоційністю. Нейромережеві моделі, такі як MOSNet і Wav2Vec, автоматизують оцінку якості мовлення, дозволяючи прогнозувати як загальну якість мовлення, так і його фонетичні та просодичні аспекти. Важливими також є метрики, які оцінюють семантичний зміст і контекст, наприклад, BLEU Score і BERTScore, що допомагають оцінити відповідність синтезованого мовлення тексту.

Список використаної літератури:

1. Frederick Jelinek. Statistical Methods for Speech Recognition. MIT Press, 1997. – p. 300–320

2. Lawrence R. Rabiner, Ronald W. Schafer. Digital Processing of Speech Signals. Prentice Hall, 1978. – p. 215–250

3. Richard M. A. Monro, Nicolas Stoll. Perceptual Evaluation of Speech Quality (PESQ): The New ITU Standard for End-to-End Speech Quality Assessment, Part I. – p. 98–112

4. Chin-Hui Lee, Frank K. Soong, Kuldip K. Paliwal. Automatic Speech and Speaker Recognition: Advanced Topics. Springer, 1996. – p. 185–210

5. Lawrence R. Rabiner, Ronald W. Schafer. Introduction to Digital Speech Processing. Now Publishers Inc, 2007. – p. 35–50

6. Francesco Camastra, Alessandro Vinciarelli. Machine Learning for Audio, Image and Video Analysis. Springer, 2015. – p. 145–170

7. Stephan Raaijmakers. Deep Learning for Natural Language Processing. Manning Publications, 2022. – p. 300–325

8. Christopher Bishop. Pattern Recognition and Machine Learning. Springer, 2006. – p. 410–440

9. Paul Taylor. Text-to-Speech Synthesis. Cambridge University Press, 2009. – p. 120–145

10. Daniel Jurafsky, James H. Martin. Speech and Language Processing (3rd Edition). Pearson, 2023. – p. 500–525

Автори статті

Іщеряков Сергій – кандидат технічних наук, доцент, Державний університет інформаційно-комунікаційних технологій, Київ, Україна.

ORCID: 0009-0007-5961-8218

Попов Антон – аспірант, Державний університет інформаційно-комунікаційних технологій, Київ, Україна.

ORCID: 0009-0006-7557-094X

Authors of the article

Ishcheriakov Serhii – Candidate of Science (technic), Associate Professor, State University of Information and Communication Technologies, Kyiv, Ukraine.

ORCID: 0009-0007-5961-8218

Popov Anton – postgraduate, State University of Information and Communication Technologies, Kyiv, Ukraine.

ORCID: 0009-0006-7557-094X