УДК 004.8

*Н. С. Кутепов**

АЛГОРИТМ ДЕКОМПОЗИЦИИ РЕЧЕВОГО СИГНАЛА НА ПРОСОДИЧЕСКИЕ И ТЕМБРАЛЬНЫЕ КОМПОНЕНТЫ В ЗАДАЧАХ ГОЛОСОВОГО КЛОНИРОВАНИЯ

Введение

Современные методы нейросетевого синтеза речи широко применяются в голосовых ассистентах, системах автоматического перевода и интеллектуальных интерфейсах. Одним из наиболее перспективных направлений является голосовое клонирование [1].

Современные методы синтеза речи основаны на архитектурах WaveNet и FastSpeech [2, 3]. Отдельное направление исследований связано с переносом просодических характеристик и управляемым синтезом речи [4].

Развитие данных технологий сопровождается не только улучшением качества синтеза, но и возникновением новых вызовов. В частности, генерация синтетической речи, близкой к естественной, может использоваться для подмены личности, несанкционированного доступа к голосовым системам и реализации атак социальной инженерии. Это делает актуальной задачу повышения интерпретируемости для обеспечения техногенной безопасности информационных систем с упором на человеко-машинное взаимодействие.

С технической точки зрения существенным ограничением современных моделей является отсутствие явного разделения составляющих речевого сигнала. Тембральные и просодические характеристики оказываются смешанными в латентном пространстве, что приводит к снижению качества управления синтезом и ухудшению межъязыковой адаптации [4].

В качестве основных методов исследования использовались методы математического моделирования, сравнительный анализ, методы цифровой обработки сигналов.

* Работа выполнена под руководством доктора технических наук, профессора кафедры «Информационные системы и защита информации» ФГБОУ ВО «ТГТУ» В. В. Алексеева.

Постановка задачи

Пусть задан речевой сигнал X, содержащий акустическую и просодическую информацию. Требуется разработать алгоритм, обеспечивающий его декомпозицию на независимые компоненты:

$$X \rightarrow (z_t, z_p),$$

где z_t – латентное представление тембра; z_p – латентное представление просодии.

К основным требованиям к алгоритму относятся:

1. независимость латентных компонент;
2. сохранение информативности представлений;
3. возможность управляемого изменения параметров синтеза;
4. устойчивость к межъязыковым различиям.

Основной сложностью при решении задачи декомпозиции является высокая степень взаимосвязи между указанными группами признаков. Изменение просодии может сопровождаться изменением спектральных характеристик сигнала, что затрудняет выделение независимых компонент и требует применения специальных методов регуляризации латентного пространства.

Реализованный алгоритм

Алгоритм основан на энкодер-декодерной архитектуре, применяемой в современных системах синтеза речи [3], и включает механизм явного разделения латентного пространства.

В отличие от существующих подходов, предложенный алгоритм предусматривает явное разбиение латентного пространства на независимые подпространства с использованием комбинированной функции регуляризации, обеспечивающей одновременное снижение корреляции и повышение устойчивости представлений.

Энкодер $E(\cdot)$отображает входной сигнал в два подпространства:

$$(z_t, z_p) = E(X).$$

Для обеспечения независимости компонент вводится регуляризационный член, минимизирующий статистическую зависимость между z_t и z_p. Функция потерь имеет следующий вид:

$$L = L_{rec} + \lambda L_{dis} + \gamma L_{stab},$$

где L_{rec} – ошибка реконструкции; L_{dis} – мера зависимости между латентными переменными; L_{stab} – член, обеспечивающий устойчивость представлений при вариации входных данных; λ, γ – коэффициенты.

Для повышения устойчивости декомпозиции используются отдельные механизмы обработки спектральных и временных характеристик сигнала.

На этапе обучения осуществляется совместная оптимизация параметров энкодера и декодера. При этом регуляризационные члены ограничивают степень смешения латентных признаков. В результате формируется структурированное латентное пространство, допускающее независимую модификацию отдельных характеристик речи.

Дополнительно в алгоритме предусматривается возможность межъязыковой адаптации. Для этого просодическое представление может модифицироваться с учетом статистических характеристик целевого языка, включая среднюю длительность фонем, особенности распределения пауз и типовые интонационные конструкции.

Декодер $D(\cdot)$восстанавливает сигнал:

$$\hat{X} = D(z_t, z_p).$$

Важной особенностью алгоритма является возможность независимого управления параметрами синтеза за счет модификации отдельных латентных компонент.

Результаты

Экспериментальная проверка алгоритма проводилась на наборе речевых данных, включающем записи нескольких дикторов с различными просодическими характеристиками. В качестве базового подхода использовалась модель с единым латентным представлением. Результаты показали, что предложенный алгоритм обеспечивает:

– более стабильное разделение тембральных и просодических характеристик;

– снижение искажений при изменении интонации;

– повышение естественности синтезируемой речи.

При варьировании вектора z_p наблюдается изменение интонационного рисунка без изменения тембра, что соответствует современным подходам к переносу просодии [4]. Аналогично, изменение z_t приводит к изменению идентичности голоса при сохранении ритмической структуры.

Дополнительно установлено, что раздельное представление признаков улучшает качество межъязыкового синтеза речи, так как позволяет адаптировать просодию к особенностям целевого языка независимо от тембра диктора [1].

С точки зрения техногенной безопасности, предложенный подход повышает интерпретируемость моделей синтеза речи. Это позволяет

использовать структуру латентных представлений для разработки методов обнаружения синтетической речи и предотвращения несанкционированного использования голосовых технологий.

Дополнительно оценивалась устойчивость модели к межъязыковому переносу. При синтезе речи на языке, отличном от языка исходного диктора, базовые методы демонстрировали ухудшение естественности речи и появление нехарактерных интонационных переходов. Использование предложенного подхода позволило снизить выраженность указанных эффектов за счет независимого управления просодическими компонентами.

Заключение

Разработан алгоритм декомпозиции речевого сигнала, обеспечивающий разделение тембральных и просодических характеристик на основе латентных представлений.

Список литературы

1. Transfer Learning from Speaker Verification to Multi-Speaker TTS / Jia Y. et al. // NeurIPS. – 2018.
2. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech / Tan X. et al. – 2020.
3. WaveNet: A Generative Model for Raw Audio / Oord A. et al. – 2016.
4. Towards End-to-End Prosody Transfer for Expressive Speech Synthesis / Skerry-Ryan R. et al. // ICML. – 2018.

Кафедра «Информационные системы и защита информации»
ФГБОУ ВО «ТГТУ»