АРХИТЕКТУРА НЕЙРОСЕТЕВОЙ СИСТЕМЫ СЕМАНТИЧЕСКОЙ ИНВЕРСИИ ДЛЯ ГЕНЕРАЦИИ ПСЕВДОРЕЧИ

Денскевич А.Д., аспирант

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Зельманский О.Б. – канд. техн. наук, доцент

Аннотация. Рассмотрена концепция семантической инверсии речи как альтернативы подавлению сигнала. Предложена архитектура нейросетевой системы, включающая модули распознавания, трансформации текста и синтеза речи с сохранением аллофонного профиля диктора. Обоснован выбор моделей для каждого этапа архитектуры (GigaAM, LLM, Fish Speech, IndexTTS-2) с приведением их количественных характеристик (WER, MOS, ELO, задержка). Показано, что реализация такой системы позволяет заменить смысл высказывания при сохранении индивидуальных голосовых признаков, что представляет собой новое направление в защите речевой информации.

Ключевые слова. Защита информации, семантическая инверсия, псевдоречь, нейросетевая архитектура, аллофонный профиль, защита речевой информации, синтез речи, сегментация речи, GigaAM, Fish Speech, IndexTTS-2, LLM.

Актуальные методы защиты речевой информации от утечки по техническим каналам основываются на подавлении сигнала – генерации акустических, виброакустических или электромагнитных помех, что маскирует речевой сигнал посредством повышения общего уровня сигнала [1, 2]. Существует некоторая нормативная база, которая регламентирует порядок защиты, к примеру: ГОСТ Р 50840-95 и СТБ ГОСТ Р 50840-2000. Однако ввиду стремительного развития цифровых технологий данная база может быть недостаточно актуальной. Актуальные методы всё ещё встречаются в защите переговорных комнат и помещений, содержащих конфиденциальную речь, но их использование со временем становится всё менее и менее надёжным из-за ряда ограничений: сами сигналы создают дискомфорт для слуха, либо могут на его влиять незаметно, но сохраняя влияние; сам сигнал помех требуется на достаточно высоком уровне, который превышает защищаемый сигнал на 6-10 дБ – такие методы могут быть в настоящее время не в достаточной степени эффективными, чтобы обеспечить надёжную защиту переговоров, так как существуют такие алгоритмы дешифровки, как шумоподавление и нейросетевое восстановление речи [3]. Нейросетевые приёмы способны адаптивно отделять полезный сигнал от лишнего сигнала, а шумоподавление является более консервативным и ручным способом, но не менее надёжным комплексом методов обработки сигналов, который может включать спектральное вычитание, непосредственное использование нейросетей, как, например, инструмент «нейросетевое подавление шума» и адаптивную фильтрацию. Такие методы позволяют выделять полезный речевой сигнал из общего сигнала, в котором могут быть помехи, что возможно даже при отрицательном отношении сигнала и шума.

Альтернативным подходом, обеспечивающим защиту речевого сигнала и конфиденциальность переговоров, может служить семантическая инверсия. В данном подходе сами слова и речь преображается в контексте смысла, выдавая по итогу нейтральную по смыслу – незаметную посторонним людям – речь, либо вовсе ложный по своему смыслу речевой сигнал. При таком подходе естественно задуматься о сохранении голосовых и эмоциональных характеристик диктора для отсутствия подозрений и внешней высокой степени естественности речи. Такой преобразованный речевой сигнал можно назвать псевдоречью, который при сохранении голосовых характеристик голоса человека, которого необходимо защитить, звучит естественно при этом сохраняя конфиденциальность информации. Такой подход не требует подавления сигнала, а обеспечивает семантическое преображение речевого сигнала в необходимый для защиты информации, хотя в то же время, имея цель запутать злоумышленника сильнее, может работать и со старыми методами. В таком случае исходная информация остаётся защищенной, а перехват сигнала и обход акустического барьера становятся бессмысленными.

Рассмотрим архитектуру нейросетевой системы семантической инверсии посредством этапов различной обработки речевого сигнала, учитывая существующие модели далее.

Выделим три последовательных функциональных блока:

1. Блок распознания и сегментации речи, обеспечивающий обработку речевого сигнала, выделяя из текста ряды транскрипций и общего аллофонного профиля диктора;
2. Блок трансформации содержания речи, который завязан на генерации семантически инвертированного текста, который будет сохранять при этом просодические средства языка, включающие эмоции, связи между словами, произношение и прочее;
3. Блок синтеза речи с управлением по аллофонному профилю, который даёт возможность создать необходимый акустический сигнал, используя имеющиеся голосовые признаки диктора.

Реализация данных блоков требует отдельные нейросетевые архитектуры, связываемые в алгоритм, где каждый этап, а точнее выбор способа прохождения этапа, нуждается в достаточной точности и работе в реальном времени, что будет определяться посредством выбора той или иной архитектуры.

Различные архитектуры и модели отрабатываются с разными языками в той или иной степени эффективности. Модель автоматического распознавания речи (ASR) наиболее эффективная для русского языка – GigaAM, разработанная SberDevices. Данная модель прошла независимые тесты, по результатам которых показатель WER (Word Error Rate – частота ошибок в словах) на центральном процессоре достигает 3,3% на русском языке. Данный показатель сравнивался с моделью Whisper large-v3-turbo, где модель от SberDevices GigaAM показала WER в 2,4 выше на графическом процессоре. Модель GigaAM обучена на 700 000 часах русскоязычной речи, которая использовала RNNT-декодер (Recurrent Neural Network Transducer – архитектура, объединяющая кодировщик, предсказатель и сеть совместной обработки для потокового распознавания) и показала пиковую задержку в 660 мс. Такая задержка достаточно близкая к отклику в реальном времени [4].

Распознавание текста с сегментацией на аллофоны происходит параллельно с описанным выше. Аллофонами являются некоторые реализации фонем в речи диктора, которые включаются в себя как позиции в слове, так и фонетическое окружение (слова до и после рассматриваемого слова). Аллофоны в различных языках отличаются своим количественным составом. Так в русском языке выделяются 22 гласных и 276 согласных вариаций аллофонов [5]. CTC-модель (Connectionist Temporal Classification – метод обучения нейросетей без пошаговой разметки, позволяющий находить границы фонем) может позволить обеспечить автоматическую сегментацию речи. Также в возможную помощь к данной модели может быть представлена мультимодальная модель SAM Audio от Meta, представленная в 2025 году. Данная модель выполняет сегментацию по текстовому запросу. Данная функция позволит выделять аллофоны строго обозначенного типа [6]. По итогу работы в данном блоке можно получить текст, состоящий из транскрипций, аллофонную последовательность вариаций сигналов по временной шкале и акустические параметры рассматриваемого текста – форманты (области спектральной концентрации акустической энергии, определяющие опознавание гласных и согласных). Кроме этого, выделяют длительность самих звуков и частоту голоса диктора. Всё это – аллофонный профиль диктора базового уровня.

Далее производим обработку речевого сигнала, представляющего себя стэк из транскрипций посредством большой языковой модели (LLM). Данная модель способна сгенерировать семантически инвертированные высказывания по заданным параметрам. Цель LLM в данном контексте – генерация текста схожей или той же тематики, изменив его смысл, скрыв конфиденциальную информацию, сохранив форму речи, описываемую просодическими характеристиками: ударения, длительности звуков, интонационный контур и другое. Таким образом, рассматривается просодические характеристики (ударение, интонация, длительности). Русскоязычные большие языковые модели более эффективно работают тогда, когда разработчик ориентирован на данную аудиторию среди пользователей данной модели – работа с языком более проработана и точна. Можно выделить некоторые открытые модели: Qwen3-32B, широко поддерживающий различные языки и DeepSeek-V3.1, который показывает хорошие результаты на русскоязычных данныъ. При использовании моделей естественно на вход подаётся широкий профиль текста, включающий стилистические особенности текста, выделение ключевых слов и других особенностей.

Третьим блоком является синтез речи по инвертированному тексту при сохранении аллофонических характеристик диктора. Такой синтез должен поддерживать управление по извлечённым акустическим признакам (conditioning). Для решения данной задачи подойдёт ряд архитектур. Fish Speech V1.5 (от fishaudio) - архитектура, которая использует DualAR технологию – авторегрессионный трансформер, показывающий акустические токены и управляющие параметры речи. Данная архитектура оценивалась рейтингом TTS Arena, набрав 1339 балла ELO, WER 3,5% для английского языка и CER 1,3% для китайского языка [7]. Данная архитектура включает в себя многозначность и может быть адаптирована для управления по внешним векторам, в том числе для описания голоса. Другая архитектура IndexTTS-2 превосходит современные zero-shot TTS модели по показателю WER, а также по сходству с диктором и эмоциональной точности [8]. Данная архитектура позволяет независимо задавать параметры голоса, но также и эмоций - учитывается весь акустический профиль диктора, что важно при изменении текста. Конкретно русский язык имеет широкое распространение в архитектуре Tacotron 2 (она адаптирована под русский язык в исследованиях СПбГУ на основе корпуса CORPRES – аннотированного корпуса русской речи, содержащего фонетическую и просодическую разметку), который имеет результат показателя MOS (Mean Opinion Score – средняя экспертная оценка) в 4,53, что достаточно близко к человеческой речи (4,58) [9]. В качестве вокодера может использоваться WaveRNN – вокодер, использование которого способно обеспечить качество сигнала на достаточно хорошем уровне, при этом не требуя больших вычислительных мощностей, в отличие от WaveNet. Таким образом, в предложенной архитектуре управляющие параметры генерируются в первом блоке в виде вектора, кодирующего аллофонный профиль диктора (формантные частоты, длительности, основная частота и другое). Этот вектор

подаётся в синтезатор как дополнительный вход, что позволяет генерировать речь, фонетически схожую с исходной речью диктора.

Реализация данной архитектуры или алгоритма архитектур требует учёта ряда сложностей. Задержка складывается из ASR (примерно 660 мс для GigaAM), генерации большой языковой модели и синтеза речи (CosyVoice2 имеет значение примерно 150 мс [10]), таким образом выходит 1-2 с, что является удовлетворительным результатом, однако с возможностью для оптимизации, посредством квантования и малых моделей. Эмоции содержаться в аллофонном профиле диктора в виде конкретной информации, которая включает просодику. IndexTTS-2 переносит эмоциональный тон исправно [8]. Большинство моделей (Fish Speech, CosyVoice2) обучены на английском и китайском; для русского языка требуется адаптация на данных типа CORPRES, что представляется возможным ввиду схожести архитектур.

Отличие маскирования речевого сигнала посредством шума и глушения сигнала отличается от семантической инверсии концептуально – в одном случае слышен шум и против его используют различные приёмы, а в ином случае можно слышать осмысленную речь, которая на самом деле не имеет конфиденциальной информации.

Таким образом была предложена трёхблочная архитектура, включающая такие технические решения, как: GigaAM, LLM-трансформация, синтез речи на Fish Speech V1.5 или IndexTTS-2 с управлением по аллофонному профилю диктора. В речи изменяется смысл при сохранении голоса - таким образом обеспечивается высокий уровень защиты информации. Однако здесь важно учитывать индивидуальные различия дикторов и возможно помещений, адаптацию моделей под русский язык и другие параметры, к которым нейросетевые технологии адаптируются в процессе обучения. Перспективными исследованиями по данной тематике представляются: прототип такой архитектуры, её оценка по различным метрикам, а также в целом лингвистическая адекватность без цифровых галлюцинаций.

Список использованных источников:

1. Зольников, В. К. [и др.]. Исследование способов защиты информации от утечки по акустическому и виброакустическому каналам / В. К. Зольников, С. А. Сазонова, А. И. Заревич, С. С. Башун // Моделирование систем и процессов. – 2025. – № 2. – С. 27–40.

2. Хорев, А. А. Техническая защита информации: учебное пособие для вузов. В 3 т. Т. 1. Технические каналы утечки информации / А. А. Хорев. – М.: Минобороны России, 2018. – 436 с.

3. Волчихина, М. В. Алгоритм формирования адаптивной речеподобной помехи для защиты конфиденциальной речевой информации в офисных помещениях / М. В. Волчихина // Труды учебных заведений связи. – 2025. – Т. 11, № 5. – С. 21–27.

4. Снижение WER с 33% до 3.3% для русской речи на CPU [Электронный ресурс]. – Режим доступа: https://www.pvsm.ru/python/445372/ – Дата доступа: 27.03.2026.

5. Бондарко, Л. В. Фонетика современного русского языка: Учебное пособие / Л. В. Бондарко. – СПб.: Издательство СПбГУ, 1998. – 276 с.

6. Bowen Shi [et al.]. SAM Audio: Segment Anything in Audio / B. Shi, A. Tjandra, J. Hoffman [et al.] // arXiv preprint arXiv:2512.18099v1. – 2025.

7. Fish-Speech-1.5 – Model Info [Электронный ресурс] // SiliconFlow. – Режим доступа: https://www.siliconflow.com/models/fishaudio-fish-speech-1-5 – Дата доступа: 27.03.2026.

8. Zhou, S. [et al.]. IndexTTS2: A Breakthrough in Emotionally Expressive and Duration-Controlled Auto-Regressive Zero-Shot Text-to-Speech / S. Zhou, Y. Zhou, Y. He [et al.] // arXiv preprint arXiv:2506.21619. – 2025.

9. Google создала систему синтеза речи, почти неотличимую от человека [Электронный ресурс] // N+1. – 2017. – Режим доступа: https://nplus1.ru/material-print/14117/ – Дата доступа: 27.03.2026.

10. FunAudioLLM/CosyVoice2-0.5B – Model Info [Электронный ресурс] // SiliconFlow. – Режим доступа: https://www.siliconflow.com/models/funaudiollm-cosyvoice2-0-5b – Дата доступа: 27.03.2026.