ПРИКЛАДНЫЕ АСПЕКТЫ ФОНЕТИЧЕСКИХ ИССЛЕДОВАНИЙ

П. А. Скрелин, А. О. Титюшина

г. Санкт-Петербург, Россия, Санкт-Петербургский государственный университет

ГЕНЕРАЦИЯ РЕЧЕВОГО СИГНАЛА С ЗАДАННЫМИ ПРОСОДИЧЕСКИМИ ХАРАКТЕРИСТИКАМИ

Данная статья посвящена разработке инструментария для моделирования эмоционального речевого высказывания с заданными просодическими характеристиками, такими, как кривые изменения частоты основного тона, интенсивности, ритмической организации и тембральной окраски. Описаны подходы к решению данной задачи, приведены результаты генерации искусственного сигнала с требуемыми характеристиками, а также данные перцептивной оценки его свойств и степени близости синтезированных сигналов к естественной речи.

Ключевые слова: просодический тембр; спектральный анализ; цифровая обработка сигнала; моделирование речевого сигнала; эмоциональная речь.

This article focuses on the development of a toolkit for modeling emotional speech utterances with specified prosodic characteristics, such as fundamental frequency contours, intensity, rhythmic organization, and timbre. The approaches to this task are described, along with the results of generating an artificial signal with the required properties and a perceptual evaluation of its naturalness parameters.

Keywords: prosodic timbre; spectral analysis; digital signal processing; speech modelling; emotional speech.

Одним из факторов, определяющих степень приближения синтезированного сигнала к естественной речи, является соответствующее коммуникативной ситуации просодическое оформление генерируемых высказываний. Наиболее сложными для реализации остаются высказывания, подразумевающие яркую эмоциональную окрашенность [1, 2]. В данном исследовании предлагаются подходы к моделированию просодических характеристик, вызывающих определенные эмоциональные ассоциации у слушателей.

В основе описываемого в данной статье инструмента лежит алгоритм ресинтеза речевого сигнала, предложенный нами ранее и описанный в статьях [3, 4]. Этот алгоритм позволяет воссоздать сигнал по периодам основной частоты, используя лишь наиболее информативные компоненты его спектра. Для достижения высокого качества результирующего сигнала достаточно использовать менее 5% рассчитанных спектральных данных. Алгоритм отбирает амплитуды и фазы частот, наиболее близких к идеальным частотам обертонов. С использованием полученных данных, строится новый синтезированный период гласного или сонорного звука. В результате 15-30 троек значений частоты, амплитуды и фазы формируют синтезированный

период основной частоты. При ресинтезе результирующий сигнал сохраняет мелодику и ритмическую организацию исходного, поскольку сохраняются длительности каждого периода частоты основного тона. Относительная интенсивность разных периодов сигнала также практически не меняется, так как для их генерации используются именно те спектральные составляющие, которые вносят в сигнал наибольший вклад. При генерации сигнала каждому периоду основной частоты задаются указанные параметры звуков, произнесенных в высказываниях с иной эмоциональной окраской.

Эффективность генерации просодического тембра была подтверждена следующим перцептивным экспериментом. Была проведена генерация 10 вопросительных и 10 утвердительных высказываний с одинаковым текстовым содержанием, произнесенных с разными эмоциями. Варианты эмоциональной окраски были следующими: уверенность, интерес, нежность, одобрение, гнев, высокомерие, угнетенность, скука, раздражение и нейтральная окраска. Каждый синтезированный сигнал содержал в себе неизменные согласные сигнала-акцептора и сгенерированные гласные и сонанты сигнала-донора тембра. Было выдвинуто предположение о том, что выделенных алгоритмом ресинтеза спектральных составляющих будет достаточно для того, чтобы аудитор смог распознать в сгенерированном на их основе сигнале исходную эмоцию сигнала-донора тембра.

В эксперименте поучаствовали 28 аудиторов (7 мужчин и 21 женщина) в возрасте от 18 до 36 лет. Для прослушивания были предложены 20 записей, 10 из которых были естественными, а 10 — сгенерированными, представляющими все приведенные выше варианты эмоциональной окраски. Вариант окраски сигнала-акцептора выбирался случайным образом. Для каждого сигнала в опроснике были заданы оригинальный и два произвольных текстовых контекста, которые могли бы окружать данную запись. Аудиторам предлагалось выбрать, какой текстовый контекст по их мнению лучше всего соответствует прослушанному сигналу.

В результате анализа полученных данных было установлено, что аудиторы правильно выбрали контекст в 77% случаев для естественных сигналов и в 70% - для сгенерированных, что позволяет сделать вывод о том, что искусственные сигналы успешно передают те же эмоциональные коннотации, что и исходные естественные. Лучше всего аудиторами были распознаны записи, передающие эмоциональные оттенки гнева (96% и 89% правильных ответов для естественного и искусственного сигналов соответственно), угнетенности (96% и 82%) и интереса (96% и 93%). Хуже всего — одобрения (50% и 32%), наглости (68% и 64%) и раздражения (71% и 75%).

Можно предположить, что выражение некоторых эмоций обеспечивается более универсальными и узнаваемыми акустическими паттернами (появление бархатности, напряженности, более светлых звуков, резких изменений тона). Эксперименты по распознаванию эмоций русскими и немецкими слушателями, в том числе и в делексикализованном сигнале [5],

также показывают значительный процент их опознания. Лексический компонент при этом играет модулирующую роль, усиливая точность распознавания, но не является обязательным условием для идентификации базовых эмоций. Исследователи также отмечают и важность акустических характеристик согласных в эмоциональной речи, что будет учтено в нашей последующей работе.

Помимо выбора текстовых контекстов для прослушанных сигналов, в конце опросника участникам перцептивного эксперимента предлагалось оценить, какой процент записей составляли искусственные. Полученные ответы можно расценивать как случайные — аудиторы оценили процент искусственных сигналов с разбросом от 0 до 60%, в среднем — 28%. Тем не менее, требуется дополнительное изучение возможных отличий между естественным и ресинтезированным сигналом, а также исследование причин возникновения возможных звуковых артефактов.

Описанный выше алгоритм позволяет не только ресинтезировать исходный сигнал, но и придавать ему новые просодические характеристики. Так, при генерации периодов, можно изменять их длительность, сдвигая на необходимую разницу значение частоты каждого обертона, но сохраняя их амплитуды и фазы, тем самым формируя необходимый мелодический контур. Однако, при значительных изменениях частоты основного тона необходимо следить за тем, чтобы зоны формантного усиления находились в пределах, характерных для синтезируемых гласных, чтобы сохранить их перцептивные свойства. Ритмическая же структура высказывания может быть смоделирована простыми способами регулирования длительности гласных или сонантов.

Для моделирования просодического тембра необходимо усиливать или ослаблять определенные спектральные составляющие сигнала. Для перехода от конкретных частот к более общим моделям распределения спектральной плотности можно воспользоваться перцептивной шкалой барков [6]. Таким образом, выделенные алгоритмом частоты можно сгруппировать по полосам барков и просуммировать их вклады, формируя некоторые модели спектральных контуров для разных эмоциональных состояний.

Для сравнения спектров аллофонов в барк-шкале были рассмотрены ударные гласные звуки [о] слова *«подобное»* фразы *«Ты где-нибудь видела что-нибудь подобное?»*, произнесенной с разными эмоциональными оттенками одним и тем же диктором. Как правило, эти гласные имели яркую тембральную окраску, интонационное движение и существенную длительность, что позволило найти среди периодов всех гласных похожие по частоте основного тона, но относящиеся к различным эмоциональным оттенкам.

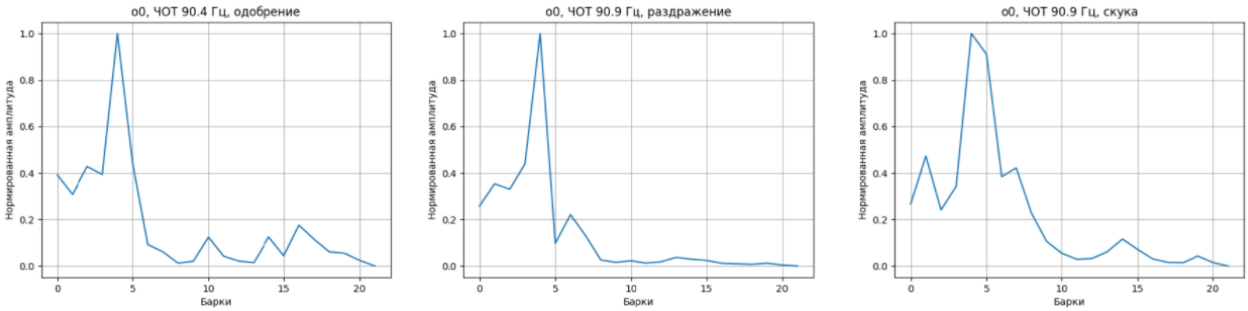

Рис.1. Сравнение спектрального профиля в шкале барков для гласного звука [о], произнесенного с частотой 90-91 Гц.+

На рис.1 можно видеть представления отобранных частот в барк-шкале для тройки эмоций «одобрение - раздражение – скука». Можно увидеть, что графики значительно отличаются как относительным вкладом первой гармоники, так и пологостью главного спектрального пика или мощностью во второй половине спектра в шкале барков. Сравнение подобных контуров позволит понять, какие именно полосы частот в сигнале необходимо усилить, а какие ослабить для достижения необходимого перцептивного отклика у слушателя.

В настоящий момент планируется построение таких моделей тембра для различных эмоциональных состояний и их применение к нейтральным и эмоционально окрашенным высказываниям, а также изучение связи проявления тембральных особенностей речи с временными и мелодическими характеристиками.

ЛИТЕРАТУРА

1. An overview of voice conversion and its challenges: From statistical modeling to deeplearning. / Sisman, B., Yamagishi, J., King, S., Li, H. - IEEE/ACM Trans. Audio Speech Lang. Process, 29, pp. 132–157 (2021).
2. Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. / Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R.J., Battenberg, E., Shor, J., Xiao, Y., Jia, Y., Ren, F., Saurous, R.A. - Proceedings of the 35th International Conference on Machine Learning, ICML. PMLR: Mc Kees Rocks, PA, USA, Volume 80, pp. 5167–5176 (2018)
3. Просодический тембр: методы анализа и модификации. / Скрелин П.А., Титюшина А.О. - Современные исследования звучащей речи, МГЛУ, Минск, 2024. с. 13-16
4. Acoustic Features of Prosodic Timbre / Skrelin P., Tityushina A. - Literature, Language and Computing; 2024 (in print)"
5. German-Russian Language Contact: Is it in our power to foresee the flight of a word we have uttered? / Skrelin, Pavel - GESUS-LINGUISTIK-TAGEN IN SANKT PETERSBURG, 23., 2015. Hamburg, 2017. p. 65-74.
6. Subdivision of Audible Frequency Rangeinto Critical Bands (Freqenzgruppen). / Zwicker, E.W. - Journal of the Acoustical Society of America, 33, 1961. p. 248.