

Фонетический лицей

Выпуск 9



300
ЛЕТ СПбГУ

Санкт-Петербург
2024

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФОНЕТИЧЕСКИЙ ЛИЦЕЙ

Выпуск 9

© Скифия-принт, 2024

ISBN 978-5-00197-131-3

ББК 81.2Рус-2
Ф77

Научные редакторы:
В. В. Евдокимова, П. А. Холявин

(6,2 Мб) **Фонетический лицей. Вып. 9:** сб. статей. — СПб.: Скифия-принт, 2024.

Данное издание является продолжением серии выпусков «Фонетического лица» — сборника студенческих работ по фонетике и речевым технологиям, который издается с 2004 года на кафедре фонетики и методики преподавания иностранных языков Санкт-Петербургского государственного университета. В текущий выпуск вошли лучшие доклады, представленные на секции «Фонетика» XXVII Открытой конференции студентов-филологов в СПбГУ, проходившей с 22 по 27 апреля 2024 года. Статьи сопровождаются аннотациями на английском языке.

Данный сборник посвящен 300-летию Санкт-Петербургского государственного университета.

Подписано к использованию 03.07.2024
ООО «Издательство Скифия-принт»
197198. С.-Петербург, ул. Б. Пушкарская, д. 10, лит. А. пом. 32-Н
Тел. (812) 982-83-94
skifia-print@mail.ru
www.skifia-print.ru/

ОГЛАВЛЕНИЕ

<i>Ананьева Анастасия Денисовна</i>	
Влияние пола на восприятие темпа речи	4
<i>Гинева Дарья Сергеевна</i>	
Восприятие интонации в синтезированной речи.....	9
<i>Гурков Иван Евгеньевич</i>	
Автоматическое распознавание и нейтрализация эмоциональной речи ..	15
<i>Гусева Дарья Дмитриевна</i>	
Устный и письменный текст сквозь призму выделения ключевых выражений	23
<i>Долгушин Михаил Дмитриевич</i>	
Распознавание эмоциональной речи людей, переживших сильное потрясение	29
<i>Ложкина Виктория Александровна</i>	
Просодические корреляты вежливости в устной русской речи. Пилотный эксперимент	35
<i>Максимова Мария Романовна</i>	
Фонетические характеристики речи при когнитивной нагрузке	44
<i>Пахомова Марина Алексеевна</i>	
Корейско-русская фонетическая интерференция	49
<i>Пискунова Жанна Евгеньевна</i>	
Русско-китайская интерференция при реализации интонации в речи на русском языке.....	58
<i>Радионова Евгения Алексеевна</i>	
Аудиовизуальное оформление прагматического маркера в общем	64
<i>Скоробагатько Лада Леонидовна</i>	
Автоматическое транскрибирование русской устной речи при помощи инструмента Kaldi.....	69
<i>Столяров Иван Игоревич</i>	
Ошибки в прочтении слов-омографов и некоторые другие факторы в восприятии синтезированной речи	75
<i>Титова Рената Руслановна</i>	
Просодические характеристики начала лекции и их восприятие.....	84
<i>Филимонова Надежда Александровна</i>	
Анимационные фильмы на занятиях по английской фонетике	90
Phonetic Lyceum, Vol. 9. Abstracts (in English).....	95

ВЛИЯНИЕ ПОЛА НА ВОСПРИЯТИЕ ТЕМПА РЕЧИ¹

Ананьева Анастасия Денисовна
(Санкт-Петербургский государственный университет)

Введение

«Что же касается восприятия, то оно является результатом психологических процессов, в которых задействованы такие понятия, как смысл, взаимосвязи, контекст, субъективная оценка, предшествующий опыт индивидуума и память» [1: 25].

Восприятие времени отличается от других перцепций, таких как восприятие цветов, пространства, движения, звуков, вкуса и др. в первую очередь, потому что «нет ни очевидных сенсорных рецепторов или органов, предназначенных для восприятия времени, ни каких-либо непосредственных, наблюдаемых ощущений, вызываемых специфическими стимулами, связанными со временем».

Восприятие темпа речи — это то, как оценивается скорость речи говорящего (им самим или собеседниками). Многогранная природа восприятия темпа речи позволяет оценивать это явление с точки зрения ряда факторов.

Человек довольно чувствителен к изменению скорости речи. Кене пишет о том, что порог различения (just noticeable difference) для темпа речи составляет 5% [5].

Ряд работ подтверждает наличие взаимосвязи пола как говорящего, так и слушающего на восприятие темпа речи. Например, в исследовании С. Филдштейна, Ф.-А. Дом и С. Краун [3] было установлено, что женщинам, говорящим быстрее, темп речи других женщин казался выше. В свою очередь, для мужчин, оказалось, что чем выше его темп, тем ниже он оценивает темп других женщин. В диссертации Моник Биemanс [2] был проведён анализ характеристик голоса, влияющих на восприятие пола говорящего. Результаты показали, что высокий голос с медленным темпом воспринимается как женский, низкий голос с высоким темпом речи воспринимается как мужской («a high, slow voice is feminine, and a low, fast voice is masculine») [2: 169].

Цель настоящего исследования — выяснить, влияет ли пол говорящего и слушающего на то, как человек воспринимает темп. В ходе работы были поставлены следующие задачи:

¹ Научный руководитель — доцент СПбГУ, канд. филол. наук У.Е. Кочеткова

- Подобрать материал для исследования.
- Подготовить аудиозаписи.
- Провести перцептивные эксперименты.
- Провести статистический анализ полученных результатов.

Материал и метод

Из корпуса CoRuss были выбраны 4 пары (мужчина и женщина) дикторов с примерно одинаковым полным темпом речи и 5 пар с близким артикуляторным темпом. Принимая во внимание тот факт, что разница по темпу в парах составляет менее 0,5 %, что является меньше порога различения [5], можно утверждать, что нет никаких объективных показателей того, что дикторы говорят с разной скоростью. Из записей были выбраны по две фразы, одинаковые для каждой пары дикторов длительностью от 2 до 13 секунд. Для исследования восприятия артикуляционного темпа были выбраны более короткие фразы, во многих из которых не содержалось пауз. В том случае, если паузы были, то их длительность в парах уравнивалась, путём добавления или удаления участка паузы. В конец каждой аудиозаписи добавлялся пустой участок для того, чтобы формальная длительность файлов в паре была одинаковой. Такое изменение было необходимым, так как тестовый эксперимент, в ходе которого участникам была доступна информация о том, что одна из аудиозаписей является длиннее, показал наличие сильного влияния на их ответ.

Опросы проводились с использованием платформы SoSci Survey. Участникам было предложено прослушать восемь пар записей в первом эксперименте, а во втором — пять пар, и определить, какая из них звучит более быстрой.

Результаты и выводы

Восприятие полного темпа

В первом эксперименте приняло участие 37 человек, 25 женщин и 12 мужчин в возрасте от 16 до 52 лет. Все ответы были сгруппированы и представлены в таблице 1. Некоторые участники ответили не на все вопросы. Оценка влияния пола участника на ответ проводилась с использованием метода хи-квадрат.

Таблица 1. Результаты первого опроса

	Мужчины	Женщины	Всего
Мужчина говорит быстрее	37	87	124
Женщина говорит быстрее	37	72	109
Затруднились ответить	15	29	44
Всего	89	188	277

Значение хи-квадрат составило 0,76, а $p = 0,31$. Из этого можно сделать вывод о том, что пол участника никак не влияет на восприятие полного темпа речи.

Тем не менее была составлена статистика ответов и посчитан хи-квадрат по каждому из вопросов отдельно. Удалось установить, что в одном из вопросов p -значение меньше критического, что говорит о значимости фактора. Лишь 8% женщин в этом вопросе затруднились дать ответ, в то время как практически половина мужчин не смогли сделать выбор. Такое распределение можно назвать необычным, так как во всех остальных вопросах в соотношении мужчин или женщин, которые затруднялись дать ответ, не наблюдается большой разницы.

Результат показывает, что мужчинам было труднее сделать выбор в данном случае. Сложно сказать, что стало причиной для этого. Важно отметить, что паузы у диктора-мужчины в этом вопросе оказались наиболее зашумлёнными (рис. 1), в некоторых местах даже присутствуют хезитации вроде «эм» (участок выделен красным на рисунке 1).

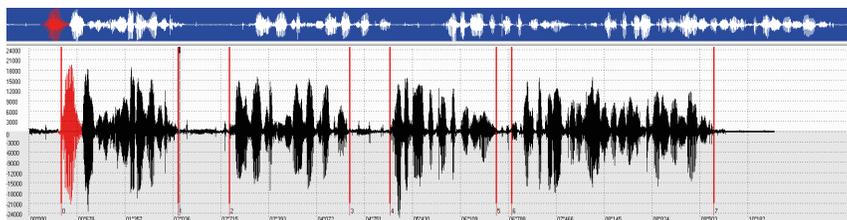


Рис. 1. Осциллограмма фразы диктора-мужчины

Вполне вероятно, что факт того, что паузы были наполнены не пустотой, а посторонними звуками, заставил участников-мужчин засомневаться в своем ответе, в то время как женщины не обратили на это внимание. Подобное наблюдение противоречит утверждению о том, что мужчины лучше справляются с различением звука в условиях шума [4]. Очевидно, что в таком случае речевой шум имеет отличную природу и, соответственно, обрабатывается другими механизмами.

Помимо прочего, идея о том, что участники склонны выбирать записи с мужскими голосами, так же не нашла подтверждения. Был проведен t-тест, выборка составила 233 ответа без учета варианта «затрудняюсь ответить». 124 раза женский голос был выбран как звучащий быстрее и 109 раз — мужской. Результаты показали, что р-значение равно 0,32, что больше, чем выбранный уровень значимости (0,05).

Восприятие артикуляторного темпа

Во втором опросе приняли участие 26 человек, 12 мужчин и 14 женщин в возрасте от 18 до 25 лет. Оценка влияния пола участника проводилась так же с использованием критерия хи-квадрат. Результаты представлены в таблице:

Таблица 2 Результаты второго опроса

	Мужчины	Женщины	Всего
Женщина говорит быстрее	38	39	77
Мужчина говорит быстрее	55	58	113
Затруднились ответить	20	28	48
Всего	113	125	238

Значение хи-квадрат составило 0,66, а $p = 0,28$, что говорит о незначимости фактора. Более того, анализ не показал значимость фактора пола участника ни в одном из отдельно взятых вопросов.

Затем была проведена оценка влияния пола диктора на выбор с помощью t-теста. Случаи, когда был выбран вариант «затрудняюсь ответить» не учитывались. Значение t-статистики составило 2,65, а $p = 0,0087$ что меньше критического.

Следовательно, значение пола диктора отказывает влияние на то, как человек воспринимает темп. Распределение ответов участников показывает, что в большем количестве случаев именно мужской голос оценивается как более быстрый: 47% всех ответов приходится на выбор записи с мужским голосом как более быстрой.

Последним шагом стал анализ ответов для мужчин и женщин отдельно. Его результаты показали, что среди женщин нет определенной тенденции к тому, чтобы они воспринимали именно женщин или мужчин быстрее ($p = 0,169$), а вот среди мужчин есть склонность считать мужские голоса более быстрыми ($p = 0,053$).

Заключение

В результате анализа данных, полученных путём проведения перцептивных экспериментов, можно сделать несколько выводов:

- внешние факторы могут оказывать значительное влияние на темп речи, например информация о длительности записи. Это важно учитывать при составлении перцептивных экспериментов, так как то, что мы воспринимаем является отражением совокупности разнообразных факторов;
- пол слушающего не влияет на восприятие ни полного, ни артикуляционного темпа;
- было обнаружено влияние заполненности пауз посторонними шумами на ответы мужчин. Кажется, что им труднее сделать выбор в таких случаях. Это может объясняться разницей в восприятии речи у мужчин и женщин;
- пол диктора не влияет на восприятие полного темпа, но влияет на восприятие артикуляционного. Вероятно, это связано с тем, что большее значение на выбор участников эксперимента по восприятию полного темпа оказывало наличие пауз, их число и длительность;
- записи дикторов-мужчин чаще оценивались как говорящие быстрее в экспериментах по восприятию артикуляционного темпа. Это подтверждает тезис Моник Биemanс [2] о том, что мужской голос оценивается как имеющий более высокую скорость.

Таким образом, можно говорить о восприятии темпа речи как о комплексном явлении, на которое оказывают влияние не только лингвистические компоненты темпа (скорость артикуляции и длительность пауз), но и социальные факторы, такие как пол, в первую очередь говорящего.

Список использованной литературы

1. Шиффман Х. Р. Ощущение и восприятие. 5-е изд. СПб.: Питер, 2003. 928 с.
2. *Biemans M.* Gender variation in voice quality. Utrecht, Netherlands: LOT, 2000.
3. *Feldstein S., Dohm F.A., Crown C.L.* Gender as a mediator in the perception of speech rate // *Bulletin of the Psychonomic Society.* 1993. Vol. 31. P. 521–524.
4. *Neff D.L., Kessler C.J., Dethlefs T.M.* Sex differences in simultaneous masking with random-frequency maskers // *The Journal of the Acoustical Society of America.* 1996. Vol. 100, no. 4. P. 2547–2550.
5. *Quené H.* On the just noticeable difference for tempo in speech // *Journal of Phonetics.* 2007. Vol. 35, no. 3. P. 353–362.

ВОСПРИЯТИЕ ИНТОНАЦИИ В СИНТЕЗИРОВАННОЙ РЕЧИ¹

Гинева Дарья Сергеевна

(Санкт-Петербургский государственный университет)

Введение

Интонационное оформление играет ключевую роль в речи, выполняя ряд важнейших функций [1; 3]:

1. Структурирование и членение речевого потока.
2. Выделение и противопоставление различных типов высказываний.
3. Обеспечение связи между элементами структуры речи и отношений между ними.
4. Выражение эмоциональных оттенков.

Просодическое оформление речевого сигнала является важной задачей синтеза речи. Однако в настоящее время интонация синтезированной речи ещё не может считаться совершенно естественной.

В настоящее время широкое распространения получают технические системы, работающие с восприятием и синтезом речи и ориентированные на общение с человеком. Из-за этого возникает необходимость сравнения синтезированной речи с естественной и оценки качества звучания первой.

В работе Санниковой Е. Г. [4] подтверждена гипотеза о том, что синтезированная речь воспринимается человеком как аналог естественной речи в шуме, в связи с чем она становится менее чётко выраженной и, соответственно, более трудной для восприятия.

Целью данного исследования является выявление особенностей интонационного оформления синтезированной речи и наиболее частотных ошибок в просодике, допускаемых синтезаторами, а также изучение восприятия синтезированной речи и перцептивная оценка естественности её звучания.

Материал и методика

Исследование построено на материале синтезированной речи. Была проанализирована речь синтезаторов «Sber Salut Speech», «Yandex

¹ Научный руководитель — доцент СПбГУ, канд. филол. наук У.Е. Кочеткова.

Cloud SpeechKit» и «Звукограм», использующих технологию Text-To-Speech. В качестве материала были использованы фонетически представительный текст «Был тихий серый вечер», 25 скороговорок, 5 стихотворений, 5 бессмысленных фраз, а также набор фраз, потенциально соответствующих каждой из 7 интонационных конструкций по классификации Е. А. Брызгуновой [Брызгунова, 1980]. Общая длительность речевого материала составила 133 минуты.

С помощью программы Praat был произведён анализ мелодических контуров всего материала с целью поиска ошибок в интонационном оформлении. Наиболее часто ошибки встречались в синтагмах с ИК-4; в то же время в собранном материале ни разу не встретилась ИК-7 и практически не встречались ИК-5 и ИК-6.

В результате сравнения мелодических контуров и акустического анализа исходного материала было обнаружено наличие нескольких существенных черт, отличающих мужские синтезированные голоса от женских. В высказываниях, произнесённых мужским голосом, наблюдалось сужение мелодического диапазона (в среднем 6.2 полутонов).

Из отобранного материала были выбраны фразы, состоящие из одной или нескольких синтагм, для проведения первого перцептивного эксперимента. Опрашиваемым было предложено оценить естественность звучания синтезированной речи по десятибалльной шкале. Чем выше балл, тем естественнее звучание. По результатам опроса было обнаружено, что мужские синтезированные голоса воспринимаются как более естественные, нежели женские (средняя оценка мужской речи составила 6.67, а женской — 5.26; медианные значения — 7.0 и 5.0 соответственно). Наиболее высокие оценки (в среднем 8.52) получил мужской голос пожилого человека, характеризующийся хриплым тембром и ещё более суженным мелодическим диапазоном по сравнению с другими мужскими голосами (в среднем 5.7 полутонов). Голос пожилой женщины также считался участниками опроса более естественным, чем прочие женские голоса (средняя оценка — 6.42). Таким образом, полученный в данном эксперименте результат свидетельствует о более естественном звучании синтезированной мужской речи по сравнению с женской.

Для второго перцептивного эксперимента была взята часть фраз, использованных в первом опросе, и их мелодические контуры были вручную модифицированы в программе Praat таким образом, чтобы они соответствовали части интонационных конструкций Брызгуновой. Были исправлены ошибки синтезаторов в выборе интонационного центра, деклинации и движении тона. Кроме того, мелодический контур стал сглажен, то есть были убраны мелодические колебания, нарушав-

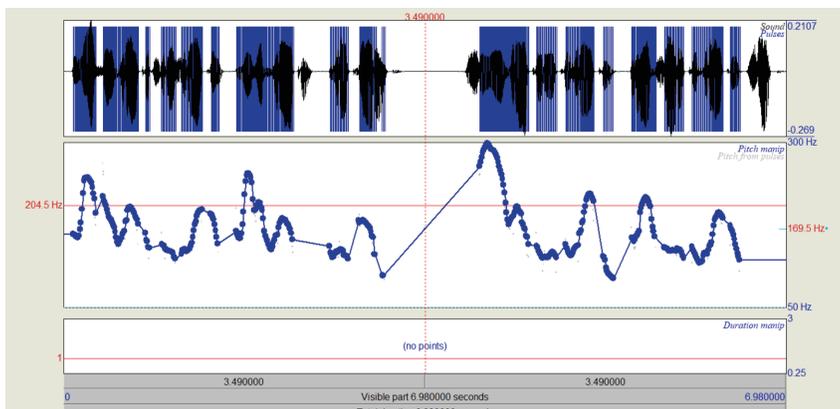


Рис. 1. Мужской голос. Мелодический контур. Скороговорка «Ехал Грека через реку»

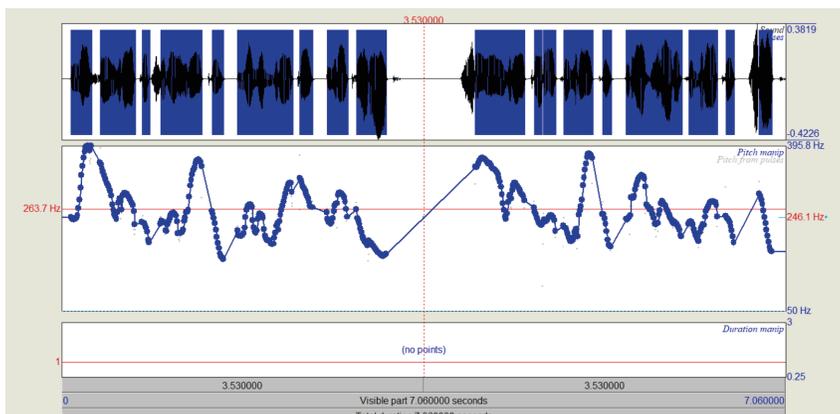


Рис. 2. Женский голос. Мелодический контур. Скороговорка «Ехал Грека через реку»

шие искомую конструкцию. При этом индивидуальная высота голоса «диктора» никак не изменялась (искусственно не понижалась и не повышалась). Для ИК-5 и ИК-6 были синтезированы специальные высказывания, которые ранее не были использованы. Получившиеся аудиозаписи были добавлены во второй перцептивный эксперимент. В этот раз опрашиваемым также было предложено оценить естественность звучания синтезированной речи по десятибалльной шкале. По результатам данного опроса было обнаружено, что оценки естественности звучащей речи вопреки ожиданиям оказались ниже, чем оценки необработанной синтезированной речи. Средние значения для женских го-

лосов составили 4,77, медианное — 4,00; для мужских — 5,90 и 6,00 соответственно. Следует отметить, что наиболее низкие оценки получил женский голос, описанный участниками эксперимента как наиболее высокий и молодой. При исключении данного голоса из статистики среднее значение для женских голосов выросло до 5,79, что выше, чем было у необработанных аудиофайлов. Наиболее высокие оценки снова заработали голоса пожилых дикторов: 7,50 (среднее) и 7,00 (медиана) для мужского голоса и 7,48 и 7,00 для женского соответственно. По результатам данного эксперимента можно сказать, что за исключением наиболее высокого женского голоса оценки естественности женских голосов стали выше, а мужских — ниже в сравнении с немодифицированными мелодическими контурами. При этом с точки зрения естественности мужские голоса стали звучать хуже, а женские — лучше.

Для третьего перцептивного эксперимента были использованы исключительно фразы, не имеющие никакого смысла, но грамматически верные с точки зрения русского языка (данные фразы были придуманы по соответствию с фразой про «Глокую куздру» Л. Щербы, которая также была включена в данный эксперимент). Как и в двух предыдущих экспериментах, аудиторам было предложено оценить естественность звучания синтезированной речи по десятибалльной шкале. По результатам данного опроса было обнаружено, что при отсутствии смысла оценки естественности изменились. Средние значения для мужских голосов составили 5,69, медианные — 6,00; для женских — 6,15 и 7,00 соответственно. Таким образом, это первый эксперимент, по результатам которого женские голоса получили более высокие оценки естественности, чем мужские. Следует также отметить, что при вычислении медианных значений для каждого голоса в отдельности, три из десяти голосов получили самые высокие результаты из трёх экспериментов (8,00). Один из них — пожилой женский голос.

Кроме того, в данном эксперименте также было посчитано, какие оценки дают аудиторы разных полов. Оказалось, что опрашиваемые женского пола в среднем оценивали любые предложенные аудиофайлы выше, чем аудиторы-мужчины (средние оценки 6,16 и 5,85 соответственно).

Отходя от мелодики, полученный синтезированный материал можно прокомментировать с точки зрения адекватности. Синтезаторы допускают большое количество таких ошибок, как постановка ударения в неправильном месте. В использованных синтезаторах заявляется, что подобная проблема может решаться постановкой специального символа (+ у SaluteSpeech, ‘ у Zvukogram), но в действительности это не всегда работает.

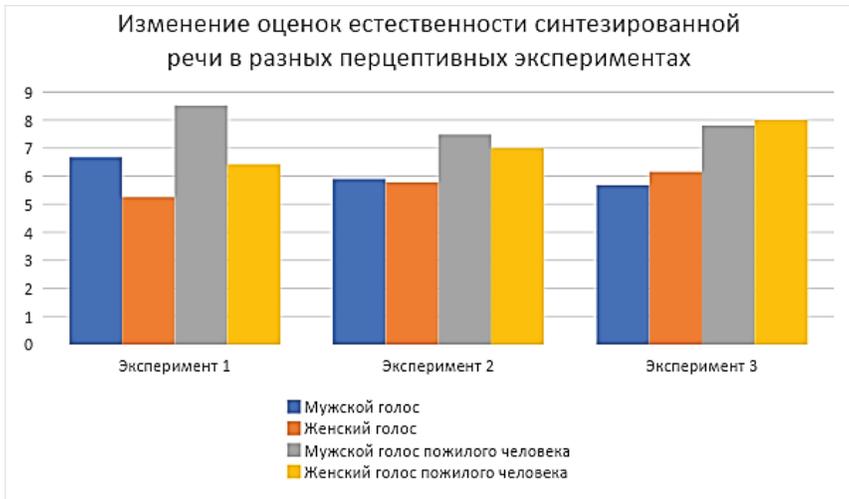


Рис. 3. Изменение оценок естественности синтезированной речи в разных перцептивных экспериментах

Синтезаторы часто допускали «странные» ошибки в чтении. Например, частица «ж» была прочитана как «же». Написанная в кавычках буква «ш» читалась как «шоссе» (при этом написанные в той же строчке «ж», «ч», «щ» были прочитаны нормально. Буква «ё» иногда читалась как «е» (Например, «всё закончилось» звучало как «все закончилось»). Буква е при этом могла читаться и как е, и как ё, причём по контексту обычно читалась правильно. Интересное явление было обнаружено при чтении е после мягких согласных, в частности [t] и [tʲ]. Вероятно, ввиду частотности е заменялся на э, а [tʲ] соответственно на твёрдый [t], несмотря на контекст. Например, в скороговорке «Булку, баранку, батон и буханку пекарь из теста испёк спозаранку» синтезатор вместо /tɛsta/ читал /testa/, то есть считал, что начальная форма слова — тест (контрольная), а не тесто. Это не исправилось и при специальной проверке другими фразами: например, в предложении «Триста грамм теста» по-прежнему звучал твёрдый [t].

Синтезаторы не всегда верно интерпретируют знаки препинания. Часто случается так, что они воспринимают запятую как точку и читают фразы перед ней с ИК-1 вместо ИК-3. Иногда бывает и наоборот. Кроме того, в большом количестве случаев они делают паузу на запятой, что в живой речи не всегда необходимо и, соответственно, звучит не очень естественно. Если в предложении встречается тире, на них тоже часто ставится неестественно длинная пауза. Также синтезаторы не могут воспроизводить эмоциональность (ремарка: у части голосов не было

выбора характеристики голоса (дружелюбный, злой и т. д.); у тех, у кого выбор был, выбирался вариант «нейтральный»). Ни один из синтезаторов не смог воспроизвести ИК-7. Предложения с восклицательным знаком в конце значительно чаще читались с ИК-1, чем с ИК-2.

Результаты и выводы

В большинстве исследованных на данный момент случаях мужские голоса звучат более естественно по сравнению с женскими. Однако ситуация меняется при удалении смысла высказывания — в таких случаях женские голоса оцениваются выше.

Голоса пожилых дикторов получают более высокие оценки естественности, чем молодые.

Список использованной литературы

1. *Бондарко Л. В.* Основы общей фонетики. СПб.: Изд-во С.-Петербур. ун-та, 2004.
2. *Брызгунова Е. А.* Интонация // Русская грамматика. Т. 1. М., 1980. С. 96–122.
3. *Светозарова Н. Д.*: Интонационная система русского языка. Л., 1982.
4. *Санникова Е. Г.* Восприятие и понимание естественной и синтезированной речи: Автореф. дис. ... канд. псих. наук. Екатеринбург, 2008.

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ И НЕЙТРАЛИЗАЦИЯ ЭМОЦИОНАЛЬНОЙ РЕЧИ¹

Гурков Иван Евгеньевич

(Санкт-Петербургский государственный университет)

Введение

В наше время важнейшую роль в коммуникации играет цифровое общение, зачастую осложнённое неоднозначностью и отсутствием живого собеседника. Особую трудность составляет неопределённость в выборе стиля общения. Сентимент-анализ (sentiment analysis) и перенос стиля (style transfer) — две процедуры обработки естественного языка (NLP), которые направлены на решение проблем такого типа. Сентимент-анализ включает в себя определение тональной полярности (положительной, отрицательной или нейтральной), эмоциональной окраски и её интенсивности в анализируемом тексте [1]. Перенос стиля, с другой стороны, предполагает изменение стилевой окраски текста при сохранении его содержания и смысла. Комбинируя эти методы, можно создать приложение, которое автоматически анализирует и преобразует текст сообщения в более нейтральную и вежливую версию, тем самым улучшая общее качество цифровой коммуникации. Целью данной работы является разработка и исследование возможностей алгоритма распознавания и нейтрализации эмоциональных выражений для помощи в подборе точных, кратких и вежливых способов выражения мысли.

Материал

В исследовании был использован набор данных, предоставленный ООО «Speech Drive»². Корпус представляет собой записи звонков в экстренные службы, сегментированные на относительно короткие отрывки от 1 до 12 секунд, а также сделанные вручную расшифровки этих записей. Общее количество отрывков — 4996. Эмоциональные катего-

¹ Научный руководитель — доцент СПбГУ, канд. филол. наук О. А. Митрофанова

² Автор выражает благодарность генеральному директору ООО «Speech Drive» канд. техн. наук В. А. Смирнову за возможность работать с набором данных и за обсуждение результатов.

рии кодируются буквой и описываются сразу несколькими близкими эмоциями:

- Г: грусть, печаль, горе, разочарование, огорчение;
- З: злость, гнев, досада;
- Р: радость, восторг, веселье, смех;
- У: удивление, изумление, растерянность, недоумение, сомнение;
- В: волнение, возбуждение, беспокойство, тревога;
- Ж: раздражение, нетерпеливость;
- Н: недовольство, возмущение;
- О: обида, расстройство;
- Б: обречённость, усталость, безысходность, слабость;
- С: страх, испуг, паника, отчаяние.

В текстах записей присутствуют особые метки:

- ? — непонятное слово («ну ? а если это не положено значит мы на улицу не имеем права выходить а какие-то люди приезжают к нам в подъезд какие-то кабели проводят»);
- % — часть слова, обрезанное слово («%градский сто один сорок три»);
- * — долгая пауза («они сейчас полицию приведут сюда * и увезут меня на волчанку умирать»);
- ъ — очень неудачный файл.

Числа расшифровываются полностью (два ноль ноль, тридцати пяти, восемьдесят пять).

Для отбора и очистки текстов были сформулированы следующие критерии:

- * (пауза) просто удаляется из текста, сам текст не отсеивается;
- символы ? и % удалять, если они являются первым или последним символом, и оставлять текст;
- в остальных случаях, отсеивать текст с данными символами;
- отсеивать текст, если он состоит полностью из чисел;
- отсеивать очень неудачные файлы (ъ);
- отсеивать очищенные тексты длиной менее 4 слов.

После проведения отбора текстов их количество уменьшилось до 3562, их распределение выглядит следующим образом:

С: 36

О: 50

Н: 848
Ж: 292
У: 211
Р: 13
Г: 70
Б: 116
З: 177
В: 1749

В связи с несбалансированностью классов было решено сгруппировать их в 4 класса:

- Грусть + Безысходность + Обида (222);
- Волнение + Страх (1750);
- Радость + Удивление (208);
- Злость + Недовольство + Раздражение (1308).

Методика

Перед переходом к методам глубокого обучения было решено провести эксперименты с моделями классического машинного обучения. Для выделения в корпусе признаков, необходимых для обучения, был выбран подход, изложенный в исследовании [2]. В работе авторы берут несколько групп типичных лингвистических признаков, проверяют внутригрупповую корреляцию для уменьшения размерности признакового пространства, и обучают на оставшихся признаках классификационную модель сложности текста. Группы признаков, использованные в статье [2]:

- 1) базовые количественные и лексические метрики;
- 2) формулы читабельности;
- 3) доли слов разных частеречных классов;
- 4) частотность лемм;
- 5) словообразование;
- 6) отдельные граммемы;
- 7) лексические и семантические признаки, неоднословные выражения, гипертекстовые связи;
- 8) синтаксические признаки;
- 9) оценки связности.

Некоторые метрики были исключены из анализа в рамках данной работы, к примеру, признак количества знаков, имеющих смысл в ана-

лизе сложности текста, но не адекватный для анализа тональности разговорной речи. Для уменьшения количества признаков строились корреляционные матрицы, с отсевом признаков по порогу корреляции 0,9 (см. рис. 1). Итоговая размерность векторов признаков равна 99.

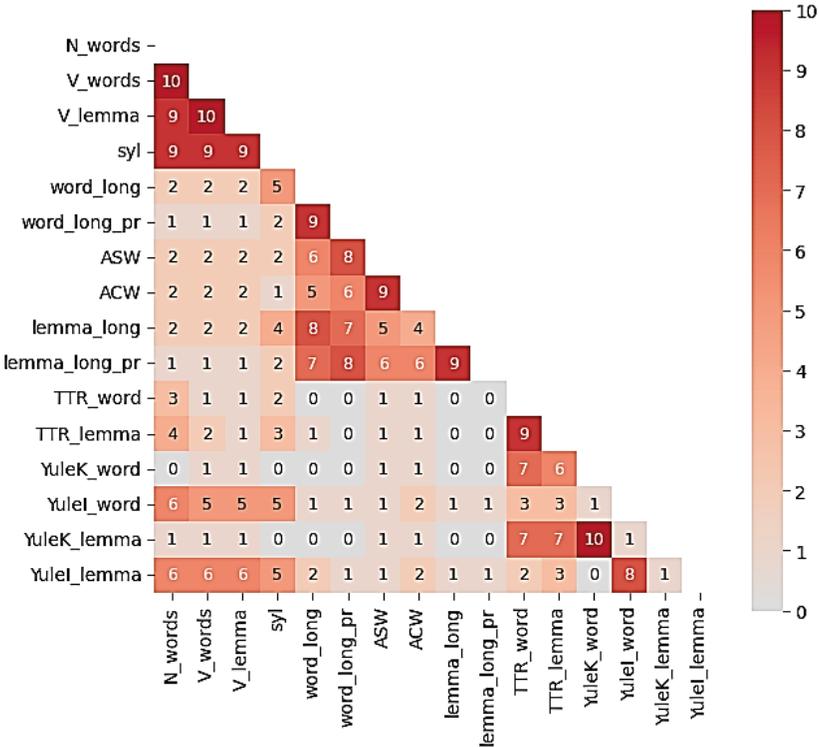


Рис. 1. Корреляционная матрица группы «Базовые признаки»

На материале полученных признаков матриц были обучены несколько классификационных моделей машинного обучения: дерево решений, случайный лес, наивный байесовский классификатор, алгоритм k-ближайших соседей, ридж-классификатор, классификация гауссовских процессов, машина опорных векторов и CatBoost. Результаты работы моделей оценивались на тестовой выборке с помощью метрик точности, полноты, аккуратности и F-меры, а также сравнения матриц ошибок и ROC-кривых. Лучшие результаты показала модель CatBoost¹ (см. табл. 1).

¹ URL: <https://catboost.ai/>

Таблица 1. Метрики модели CatBoost, обученной на лингвистических признаках

MSE	20.7
Accuracy	0.49
F-score	0.45
ROC-AUC	0.63

Также был проведён эксперимент с обучением классификационных моделей с использованием эмбедингов предложений, полученных из предобученной языковой модели Sentence RuBERT¹ от DeepPavlov, в качестве признаков. При использовании эмбедингов лучшие результаты были получены при обучении машины опорных векторов (см. табл. 2).

Таблица 2. Метрики модели, обученной на эмбедингах RuBERT

MSE	17.65
Accuracy	0.49
F-score	0.41
ROC-AUC	0.68

Следующим шагом стала проверка классификации высказываний с помощью акустической информации. Классическим методом кодирования звучащей речи является вычисление мел-кепстральных коэффициентов (MFCCs) [3], позволяющих представить спектральную информацию в сжатом и в то же время приближенном к человеческому восприятию виде. При обучении с использованием 13 коэффициентов был получен следующий результат (см. табл. 3).

Таблица 3. Метрики модели, обученной на MFCCs

MSE	15.1
Accuracy	0.62
F-score	0.61
ROC-AUC	0.73

¹ URL: <https://huggingface.co/DeepPavlov/rubert-base-cased-sentence>

Лучший результат в задаче классификации был достигнут при одновременном использовании акустических и лингвистических признаков (см. табл. 4).

Таблица 4. Метрики модели, обученной на MFCCs и лингвистических признаках

MSE	14.6
Accuracy	0.62
F-score	0.60
ROC-AUC	0.78

Для обучения модели перефразирования эмоционально нагруженной речи необходимо подготовить параллельный набор данных, представляющий из себя пары «эмоциональная речь ~ нейтральная речь», в которых сохраняется смысловое содержание, но различна эмоциональная окраска. Для формирования подобного набора данных было решено использовать методику обучения «учитель — ученик». В данном виде обучения используются две модели, одна из которых уже способна выполнять поставленную задачу, а вторая должна «учиться на её примере». Подобная методика используется, например, при обновлении архитектуры модели, когда новая модель учится у старой. В нашем случае модель с небольшим количеством параметров и узкой специализацией на задаче нейтрализации эмоций будет учиться у крупной широко направленной модели.

В качестве учителя была выбрана модель GPT-4¹ от OpenAI [4], на данный момент являющаяся SOTA (State of the Art) практически во всех задачах языкового моделирования. Недостатком этой модели является её размер и закрытость: веса и архитектура данной модели не выложены в открытый доступ, а для её работы необходимы крайне большие вычислительные мощности. С помощью обращения к модели через API был сформирован набор нейтрализованных версий высказываний из изначального набора данных.

В качестве ученика использовалась предобученная русскоязычная sequence2sequence модель ruT5-base² от SberDevices [5]. Данная модель архитектурно соответствует оригинальной модели T5³ от Google [6], но обучена с нуля полностью на русскоязычном материале (300 ГБ тек-

¹ URL: <https://openai.com/index/gpt-4/>

² URL: <https://huggingface.co/ai-forever/ruT5-base>

³ URL: <https://research.google/blog/exploring-transfer-learning-with-t5-the-text-to-text-transfer-transformer/>

стовых данных). Дообучение проводилось на вычислительных мощностях кафедры фонетики СПбГУ.

В качестве метрики качества рефразирования использовалась метрика ParaScore¹ [7]. Основная идея ParaScore заключается в том, чтобы явно учитывать два важных критерия хорошей парафразы: семантическую близость к исходному предложению и лексическое отличие от него. ParaScore комбинирует достоинства ориентированных на референс (reference-based) и не ориентированных на референс (reference-free) метрик, а также явно моделирует лексическое различие. В экспериментах ParaScore показала лучшую корреляцию с человеческими оценками по сравнению с популярными метриками, такими как BLEU, ROUGE, BERTScore и др.

В ходе экспериментов было выявлено, что модель начинает переобучаться примерно после восьмой эпохи обучения: в то время как функция потерь на обучающей выборке продолжает медленно убывать, функция потерь на валидационной выборке начинает расти. Нередко случается, что при росте функции потерь при валидации целевая метрика продолжает улучшаться, но в нашем случае это не так: примерно после той же эпохи значение ParaScore также начинает идти вниз. При финальном обучении модель достигла значения 0,826 в метрике ParaScore. Модель доступна в репозитории Hugging Face².

Выводы

Результаты данной работы демонстрируют перспективность применения методов машинного обучения и обработки естественного языка для задач распознавания эмоциональной окраски речи и ее нейтрализации. Наилучшие результаты классификации эмоций (62 % точности) были достигнуты при совместном использовании лингвистических признаков и акустической информации в виде мел-кепстральных коэффициентов.

Комбинирование крупной языковой модели GPT-4 в качестве «учителя» и более компактной модели ruT5 в качестве «ученика» позволило успешно обучить модель автоматическому рефразированию эмоциональных высказываний в нейтральную форму. Достигнутое значение метрики ParaScore 0.826 свидетельствует о способности обученной модели сохранять фактическое содержание исходных текстов при нейтрализации их эмоциональной окраски.

¹ URL: <https://github.com/shadowkiller33/ParaScore>

² URL: <https://huggingface.co/ZovutVanya/ruT5-EmotionNeutralization-MCHS>

Заключение

В данной работе были изучены и применены современные методы обработки естественного языка, такие как сентиментный анализ и перенос стиля, для решения задачи автоматического распознавания и нейтрализации эмоциональной речи. Полученные результаты демонстрируют потенциал использования данных подходов для улучшения качества цифровой коммуникации путем преобразования эмоционально окрашенного текста в более нейтральную форму при сохранении смыслового содержания. Дальнейшие исследования могут быть направлены на улучшение точности классификации эмоциональной речи, а также на расширение и дообучение модели перефразирования на больших наборах данных для повышения качества нейтрализации различных типов эмоциональной речи.

Список использованной литературы

1. Лукашевич Н. В. Автоматические методы анализа тональности // Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие. Изд-во НИУ ВШЭ, 2017. С. 127–194.
2. Блинова О., Тарасов Н. Метрики сложности русских правовых текстов: отбор, использование, первичная оценка эффективности // Компьютерная лингвистика и интеллектуальные технологии: по материалам международной конференции «Диалог 2022». Российский государственный гуманитарный университет, 2022. С. 1017–1028.
3. Mermelstein P. Distance measures for speech recognition, psychological and instrumental // Pattern recognition and artificial intelligence. 1976. Vol. 116. P. 374–388.
4. Open A. I. et al. GPT-4 Technical Report: arXiv:2303.08774. arXiv, 2024.
5. Zmitrovich D., Abramov A., Kalmykov A. et al. A Family of Pretrained Transformer Language Models for Russian: arXiv:2309.10931. arXiv, 2024.
6. Raffel C., Shazeer N., Roberts A., Lee K., Narang Sh., Matena M, Yanqi Zhou, Wei Li, Liu P. J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer: arXiv:1910.10683. arXiv, 2023.
7. Shen L., Liu L., Jiang H., Shi S. On the evaluation metrics for paraphrase generation // Proceedings of the 2022 conference on empirical methods in natural language processing. 2022. P. 3178–3190.

УСТНЫЙ И ПИСЬМЕННЫЙ ТЕКСТ СКВОЗЬ ПРИЗМУ ВЫДЕЛЕНИЯ КЛЮЧЕВЫХ ВЫРАЖЕНИЙ¹

Гусева Дарья Дмитриевна

(Санкт-Петербургский государственный университет)

Введение

В современной лингвистике принято представление о коммуникации как о мультимодальном многоуровневом процессе, использующем разные способы выражения, из которых речевая коммуникация обеспечивает вербальное взаимодействие людей. При такого рода коммуникации передача информации может осуществляться посредством как устной, так и письменной речи, которые не считаются равноправными. Большое количество исследований посвящено изучению производства и восприятия естественной речи, как устной, так и письменной [2, 4]. Установлено, что механизмы восприятия письменного и устного текста различны.

Восприятие письменного текста начинается с визуального распознавания графических символов, восприятие устного текста — со звукового восприятия, когда высказывание, реализованное говорящим, преобразуются им в звуковые колебания и затем воспринимается органами слуха [1].

Восприятие письменного текста в основном представляет собой нелинейное предсказание: читающие одновременно видят слова, окружающие рассматриваемое слово, то есть имеют доступ к левому (в большей степени) и правому контекстам. Однако при восприятии устного текста предсказание обычно линейно: доступен практически лишь левый контекст [6].

Восприятие содержания письменного и устного текста различается в том числе на уровне выделения ключевых выражений, которые способствуют пониманию смысла текста как целостного объекта. Считается, что в устном тексте ключевые выражения также обладают определённой просодической специфичностью [5].

Данное исследование нацелено на сопоставление различий в восприятии одного и того же текста, представленного в письменной и устной формах.

¹ Научный руководитель — доцент СПбГУ, канд. филол. наук О. А. Митрофанова

Актуальность обусловлена необходимостью учёта стратегий выделения ключевых выражений в устных и письменных текстах в задачах, предполагающих автоматизацию этого процесса, в частности, при индексировании текстов, при генерации транскриптов и аннотаций в веб-сервисах для семантической обработки текстов.

Материал

В качестве материала для исследования использованы аудиозаписи выступлений русскоязычных лекторов проекта «Постнаука» [3]. Каждая аудиозапись сопровождается текстовым транскриптом. Для исследования отобраны аудиозаписи двух дикторов женского и мужского пола, которые являются профессиональными лекторами и работают с разновозрастной аудиторией. Лекции принадлежат к научно-популярному стилю и рассчитаны на неподготовленную аудиторию, что позволяет привлекать к участию в опросах аудиторов без учёта их профессиональной специализации.

Метод

Наборы ключевых выражений были получены в результате серии перцептивных экспериментов с привлечением аудиторов. Аудиторы должны были после прочтения или прослушивания текста выделить из него 10 ключевых выражений, ранжируя их от самого важного к менее важному. В инструкции было установлено ограничение на тип выражений — аудиторы могли выбрать уни-, би- или триграммы (ключевые выражения, состоящие из одного, двух или трёх слов соответственно).

Были отмечены различия в восприятии текстов аудиторами, относившимися к разным возрастным группам. Чтобы проверить эту гипотезу, результаты, полученные по каждому тексту, были разделены на две группы по возрастному критерию. Участники старше 18 лет были отнесены к группе взрослых, а участники в возрасте 18 лет и младше — к группе подростков.

В эксперименте по письменным текстам к группе взрослых были отнесены 45 читателей, к группе подростков — 21 читатель. В эксперименте по первому устному тексту приняло участие 27 взрослых и 22 подростка, по второму устному тексту — 28 взрослых и 21 подросток.

Результаты и выводы

В эксперименте по первому письменному тексту взрослые выделили 64 уникальных ключевых выражения. На 10 самых частотных ключевых выражений пришлось 273 ответа (60% от общего числа ответов, данных взрослыми). Подростки выделили 105 уникальных выражений, на первые 10 самых частотных пришлось 83 ответа (39% от всех ответов, данных подростками). Соотношение между числом ответов, которые получили 10 первых по частотности ключевых выражений, и общим числом выделенных выражений для подростков ниже, чем для взрослых. Этот показатель позволяет говорить о большей вариативности в наборе ключевых выражений, полученном от группы подростков. Кроме того, в наборе, выделенном взрослыми, 5 выражений имели от-

Таблица 1. Сопоставление результатов выделения ключевых выражений взрослыми и подростками в первом письменном тексте

Взрослые			Подростки		
Ключевое выражение	Частотность	Ранг	Ключевое выражение	Частотность	Ранг
Универсальная грамматика	0.091	64	Когнитивный взрыв	0.090	105
Когнитивный взрыв	0.089	63	Грамматические гены	0.062	104
Грамматический взрыв	0.087	62	Грамматический взрыв	0.057	102.5
Грамматические гены	0.080	61	Универсальная грамматика	0.057	102.5
Структурные закономерности	0.060	60	Мнемотехника	0.029	101
Ноам Хомский	0.049	59	Эксперимент	0.024	100
Язык	0.042	58	Грамматика	0.019	96.5
Устройство овладения языком	0.040	57	Словарный запас	0.019	96.5
Ищи структуру в хаосе	0.036	56	Стремление усматривать структуру	0.019	96.5
Языковые высказывания	0.033	54.5	Итерированное обучение	0.019	96.5

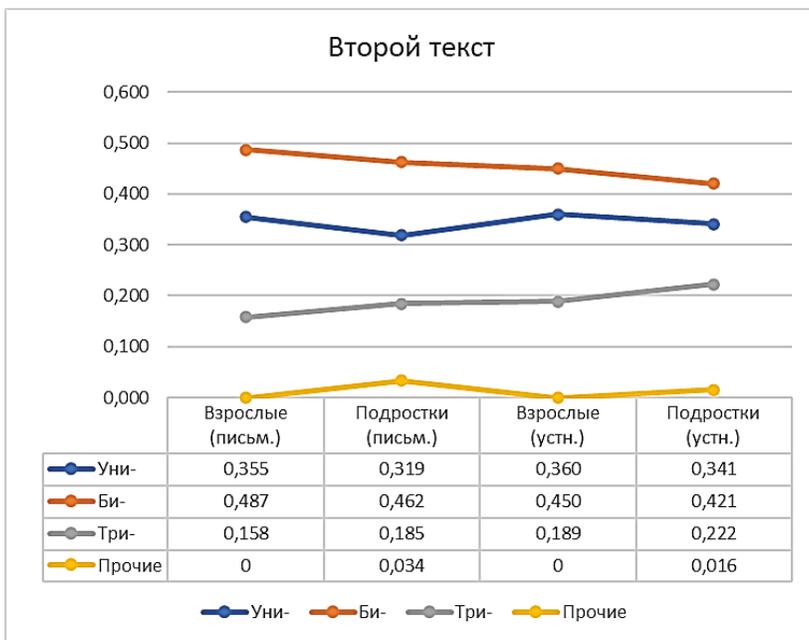


Рис. 1. Распределение ключевых выражений в ответах взрослых и подростков по второму тексту («уни-» — униграммы, «би-» — биграммы, «три-» — триграммы, прочие — элементы, содержавшие иное количество слов)

носительную частотность выше 0.05. В наборе, выделенном подростками, таких выражений было 4, однако значения частотности для них ниже, чем значения для совпадающих выражений из набора взрослых (см., например, ключевое выражение «универсальная грамматика»).

В табл. 1 представлены наборы самых частотных ключевых выражений, выделенных взрослыми и подростками в первом письменном тексте.

Аналогичным образом были проанализированы ответы, полученные по остальным текстам. В случае со всеми текстами соотношение между числом ответов, которые получили самые частотные ключевые выражения, и общим числом ответов ниже для подростков, чем для взрослых. По устным текстам для обеих возрастных групп были получены более низкие значения, чем по письменным.

Было также рассмотрено, какого типа ключевые выражения предпочитали выделять аудиторы. Результаты по второму тексту показывали достаточно однозначные закономерности (рис. 1). Для аудиторов обеих возрастных групп более характерным оказалось выделением биграмм. Биграммы во всех группах получили более 40% от общего

числа ответов, униграммы — более 30%, триграммы — более 15%. Для первого текста закономерности менее выражены, но биграммы во всех группах, кроме одной, получили более 45% ответов. Это может свидетельствовать о том, что вне зависимости от возраста аудиторы больше предрасположены к выбору ключевых выражений, состоящих из двух элементов.

Кроме того, была проанализирована частеречная принадлежность ключевых выражений, выделенных аудиторками. Однако анализ не выявил статистически значимой разницы в использовании частей речи взрослыми и подростками.

Заключение

На основании проведённого анализа можно сделать вывод, что восприятие текстов различается у аудиторов разных возрастных групп. Взрослым свойственна большая согласованность при выделении ключевых выражений, чем подросткам. Наборы уникальных ключевых выражений, выделенные подростками, во всех случаях больше, чем наборы, выделенные взрослыми. Кроме того, относительная частотность ключевых выражений выше для взрослых: в их ответах больше выражений имеют частотность, превышающую 0,05. Это может указывать на то, что взрослые более избирательно выбирают ключевые выражения и придают большее значение определённым понятиям в тексте. Данные выводы подчеркивают важность учёта возрастных различий при анализе восприятия текстов и могут быть полезны для разработки образовательных программ и методик обучения, учитывающих возрастные особенности аудитории.

Список использованной литературы

1. *Кодзасов С. В., Кривнова О. Ф.* Общая фонетика. М., 2001.
2. *Леонтьев А. А.* Основы психолингвистики. М.: Academia. 2005. 287 с.
3. Постнаука [Электронный ресурс]. URL: <https://postnauka.ru/>
4. *Сахарный Л. В., Штерн А. С.* Набор ключевых слов как тип текста // Лексические аспекты в системе профессионально-ориентированного обучения иноязычной речевой деятельности. Пермь: Пермский политехнический ун-т, 1988. С. 34–51.
5. *Ягунова Е. В.* Роль ключевых слов при восприятии звучащего и письменного текста (на материале русского языка) // Человек пишущий и читающий: проблемы и наблюдения: Материалы и наблюдения: Материалы международной конференции 14–16 марта 2002 г. Санкт-Петербург. СПб.: Изд-во С.-Петербурга, 2004. С. 197–204.

6. Ягунова Е. В. Контекстная предсказуемость и коммуникативная структура в процедурах восприятия звучащего текста // Фонетика и нефонетика. К 70-летию Сандро В. Кодзасова. М.: Языки славянских культур. 2008а. С. 630–641.

РАСПОЗНАВАНИЕ ЭМОЦИОНАЛЬНОЙ РЕЧИ ЛЮДЕЙ, ПЕРЕЖИВШИХ СИЛЬНОЕ ПОТРАСЕНИЕ¹

*Долгушин Михаил Дмитриевич
(Санкт-Петербургский государственный университет)*

Введение

С 2001 по 2006 год чешские и американские специалисты в сотрудничестве с фондом the Shoah Visual History Foundation (VHF) [8] решали задачи проекта Multilingual Access to Large Spoken Archives (MALACH) [6]. Основной целью данного проекта было улучшение методов доступа к многоязычным архивам устной речи, собранным VHF. Задача же состояла в том, чтобы развить и определить техники автоматического распознавания для спонтанной, эмоциональной и имеющей сильный акцент речи людей, переживших Холокост. Результаты проекта в дальнейшем использовались для аннотирования архивных материалов, поиска ключевых слов и т. д. Результирующая точность по метрике WER (Word Error Rate) [9] распознавания речи для чешского и русского составила 38.57 % и 45.75 % соответственно [7], чего было достаточно для выделения ключевых слов и аннотирования корпуса.

Некоторое время назад данная задача была поставлена вновь, и новейшие технологии распознавания речи позволили достичь значительных успехов для английского и чешского языков [5], однако аналогичная работа на русском языке не ведется.

Из вышесказанного следует задача автоматического распознавания эмоциональной, имеющей сильный акцент речи на русском языке свидетелей Холокоста с применением современных нейросетевых технологий.

Решение данной проблемы упростило бы автоматическое создание субтитров к доступным архивным записям по данной теме. Также данная задача представляет интерес технического характера в силу необходимости учета речевых особенностей пожилых интервьюируемых людей и того, что многие из них используют диалектизмы в речи.

¹ Научный руководитель — доцент СПбГУ, канд. филол. наук В. В. Евдокимова.

Материал

Материалом послужили видеозаписи интервью со свидетелями Холокоста, записанные фондом Яд ва-Шем [10], доступные в YouTube и имеющие субтитры. Данный набор состоит из 25 видео длиной от 40 минут до 2 часов 30 минут. Суммарная длина — 32 часа 57 минут 17 секунд. Видео данного набора также содержат речь двух человек: интервьюера и интервьюируемого. Видеозаписи могут содержать шумы, осложняющие распознавание речи. Также в видео присутствуют слова и топонимы на иврите, идише. Данные записи составили основу для формирования корпуса RuOH.

Ранее в статье [3] были описаны сбор набора RuOH и его преобразование, в результате которой суммарная длительность материала сократилась до 17,5 часов, где 7,5 часов — записи с интервьюируемыми мужчинами и 10 часов — записи с интервьюируемыми женщинами. В данной работе использовалась версия набора, опубликованная на Hugging Face [4] и разделенная на три выборки с непересекающимися интервьюируемыми. Набор имеет только орфографические транскрипции.

Метод

В качестве метода распознавания речи было решено использовать глубокую нейросетевую модель Wav2Vec 2.0, дообученную на русском языке [1] и дополнительно обучить её на выбранном материале.

Также решено изучить влияние добавления языковой биграммной модели на качество распознавания.

Поскольку в записях может быть одновременно два говорящих: интервьюер и интервьюируемый, то решено использовать нейросетевую модель ruannotate [2] для автоматического определения говорящих. В результате выяснилось, что почти 30 % записей содержат речь сразу двух говорящих. Поэтому дополнительно рассмотрено качество распознавания на записях одновременно двух говорящих.

Для определения качества распознавания используются метрики: коэффициент неверно распознанных слов (word error rate, WER) и коэффициент неверно распознанных символов (character error rate, CER).

Коэффициент WER используется для сравнения распознанных слов с действительно произнесенными словами. Учитываются три типа ошибок: замена одного слова другим, удаление слова, вставка слова. WER рассчитывается следующим образом:

$$\text{WER} = \frac{I + D + S}{N} \cdot 100\%,$$

где D , I , и S — количество неверно удаленных, вставленных и замененных слов соответственно, N — общее число слов в распознаваемом сообщении.

Коэффициент CER рассчитывается аналогично WER для распознанных символов.

Результаты и выводы

В таблице 1 представлены результаты проверки точности модели Wav2Vec 2.0 до дополнительного обучения на материале RuOH и после обучения, а также с использованием и без использования биграммной языковой модели (ЯМ) на подвыборках тестовой выборки с одним и двумя дикторами на записи.

Из таблицы можно заметить, что настройка на данных RuOH дала прирост в ~8 % WER на одном дикторе и ~10 % WER на двух дикторах. Качество распознавания речи на двух дикторах ожидаемо ниже, чем на одном дикторе, в силу трудности распознавания при наложении двух речевых потоков.

Таблица 1. Оценка качества моделей на записях с одним и двумя дикторами

Выборка	Количество дикторов	Метрика	Модель			
			wav2vec2 обученная на русском	wav2vec2 обученная на русском + настройка на RuOH	wav2vec2 обученная на русском + ЯМ	wav2vec2 обученная на русском + настройка на RuOH + ЯМ
Тестовая	1 диктор	WER (%)	48,45	40,62	41,22	35,88
		CER (%)	25,60	22,58	24,58	23,26
	2 диктора	WER (%)	54,59	44,49	46,70	39,89
		CER (%)	31,11	26,96	30,67	27,51

Также значительный прирост в качестве дало добавление языковой модели, позволившее снизить WER до 35,88% на одном дикторе и 39,89% на двух дикторах. Хотя по метрике CER качество модели оказалось выше при использовании только акустической модели, и достигло 22,58%. Возможно, такое расхождение может быть связано со случаями, когда ошибочно одно слово разделяется на два, в результате чего языковая модель исправляет обе части до корректных орфографически

слов русского языка, чем создает лишние знаки, значительно влияющие на коэффициент CER, но дающие незначительный вклад в коэффициент WER.

Далее был проведен лингвистический анализ встречающихся расхождений распознанных текстов с субтитрами для сравнения.

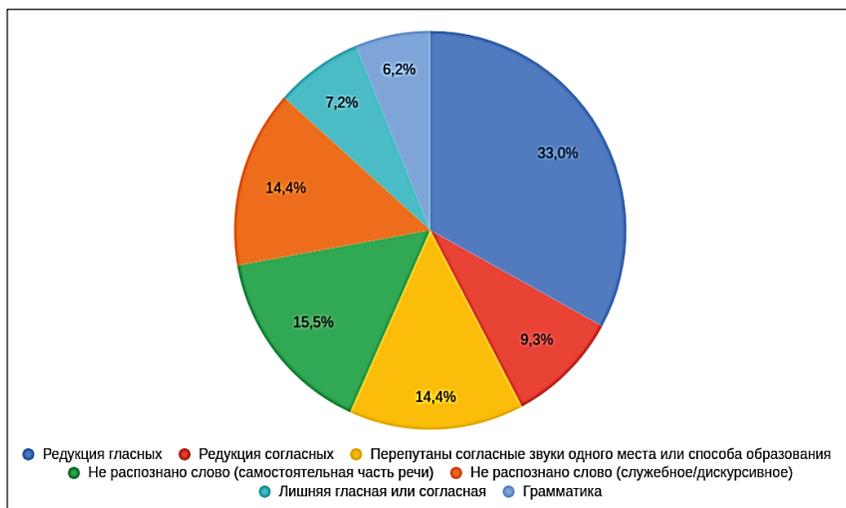


Рис. 1. Анализ несоответствий расшифровкам нейросети без добавления языковой модели

Наибольший процент ошибок (33%) приходится на проблемы с распознаванием гласных звуков из-за сильной редукции в спонтанной речи. В том числе в данный пункт включены ошибки, вызванные расхождением орфографической записи и произносительной нормой некоторых гласных.

Примеры возникающих ошибок:

- замена «о» на «а» и наоборот в безударных позициях. Замена «а» на «о» может быть объяснена украинско-русской интерференцией в речи одного из дикторов в тестовой выборке;
- замена «е» на «и» в конце слова;
- замена окончания прилагательных «ая» на «ое».

Несмотря на высокую частотность данного рода ошибки, многие из них исправляются добавлением языковой модели, помогающей учитывать грамматику и орфографию русского языка.

Замечено, что в некоторых случаях модель, благодаря дообучению, справляется со ошибками, возникающими из-за интерференции, одна-

ко в низкочастотных словах качество становится хуже. Для решения данного недостатка можно использовать языковую модель.

По результатам проведенного исследования можно сделать следующие выводы:

- Эксперименты с использованием интегральной акустической модели Wav2Vec 2.0 показали, что дополнительное обучение нейросетевых моделей даже на относительно небольшом наборе данных, отягощенных затрудняющими факторами, такими как эмоциональное состояние человека, возраст, высокая степень спонтанности и интерферированности речи, позволяет увеличить качество распознавания речи.
- Еще большее повышение качества распознавания может быть достигнуто при добавлении языковой модели, исправляющей орфографические ошибки в полученных расшифровках текста, возникающие из-за высокой степени межъязыковой интерференции и спонтанности речи.

Заключение

В данной работе было проведено исследование методов распознавания речи людей, переживших сильное эмоциональное потрясение, на основе интервью с людьми, пережившими Холокост, записанными фондом мемориального комплекса истории Холокоста Яд ва-Шем.

В результате, дополнительное обучение модели на основе архитектуры Wav2Vec 2.0, позволило создать дикторонезависимую систему распознавания речи людей, переживших сильное эмоциональное потрясение. Качество распознавания созданной системы в среднем на 8 % по метрике WER превосходит результаты полученные моделью без дополнительного обучения и без языковой модели. Добавление языковой модели позволило снизить WER еще на 6 %.

Также замечено снижение качества распознавания на записях с двумя говорящими и при слишком сильной степени интерференции, но результаты все еще превосходят использование модели без дополнительного обучения.

Наивысшее качество распознавания в 35,88 % WER достигается на записях с одним диктором.

В будущем предполагается создание интерфейса поиска по записям корпуса, использующее полученную модель для определения более точных временных границ в видео, в которых находится ответ на во-

прос, а также для добавления в корпус видеозаписей фонда Яд ва-Шем, к которым нет подготовленных субтитров.

Список использованной литературы

1. *Bondarenko I.* XLSR Wav2Vec2 Russian by Ivan Bondarenko [site]. URL: <https://huggingface.co/bond005/wav2vec2-large-ru-golos> (date of application: 18.07.2022).
2. *Bredin H.* Pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. DOI: 10.21437 / Interspeech.2023-105 // Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2023. ISCA, 2021. P.1983–1987. URL: https://www.isca-archive.org/interspeech_2023/bredin23_interspeech.pdf (date of application 26.04.2024).
3. *Bukreeva L.* Emotional Speech Recognition of Holocaust Survivors with Deep Neural Network Models for Russian Language / L. Bukreeva, D. Guseva, M. Dolgushin, V. Evdokimova, V. Obotnina // Karpov A., Samudravijaya K., Deepak K. T., Hegde R. M., Agrawal S. S., Prasanna S. R. M. (eds). Speech and Computer. SPECOM 2023. Lecture Notes in Computer Science. Vol. 14338. Springer, Cham, 2023. https://doi.org/10.1007/978-3-031-48309-7_6.
4. Demonstration version of RuOH data on Hugging Face [site]. URL: https://huggingface.co/datasets/Mihaj/ruoh_demo (date of application 12.12.2023).
5. *Lehečka J., Pšutka J. V., Pšutka J.* Transformer-based Automatic Speech Recognition of Formal and Colloquial Czech in MALACH Project. 2022. <https://doi.org/10.48550/arXiv.2206.07666>.
6. Project MALACH // MALACH — Multilingual Access to Large Spoken Archives [site]. URL: <https://malach.umiacs.umd.edu/> (date of application: 15.07.2022)
7. *Pšutka J.* Automatic transcription of Czech, Russian and Slovak spontaneous speech in the MALACH project / J. Pšutka, P. Ircing, J. V. Pšutka, J. Hajič, W. Byrne, J. Mírovský // Eurospeech. 2005. P.1349–1352. <http://dx.doi.org/10.21437/Interspeech.2005-489>.
8. USC Shoah Foundation: noncommercial foundation [site]. URL: <https://sfi.usc.edu/> (date of application: 18.07.2022).
9. *Wang Y., Acero A., Chelba C.* Is Word Error Rate a Good Indicator for Spoken Language Understanding Accuracy // IEEE Workshop on Automatic Speech Recognition and Understanding: St. Thomas: US Virgin Islands: IEEE, 2003. URL: <https://ieeexplore.ieee.org/document/1318504> (date of application: 18.07.2022).
10. Видеосвидетельства переживших холокост фонда Яд Вашем // YouTube. com = Ютьюб [сайт]. URL: <https://www.youtube.com/playlist?list=PLanQ0T-FmIYBTv8sRAkSDWQLZNhbM-v1xp>

ПРОСОДИЧЕСКИЕ КОРРЕЛЯТЫ ВЕЖЛИВОСТИ В УСТНОЙ РУССКОЙ РЕЧИ. ПИЛОТНЫЙ ЭКСПЕРИМЕНТ¹

Ложкина Виктория Александровна
(Санкт-Петербургский государственный университет)

Введение

Вежливость регулирует межличностные отношения, помогает поддерживать социальную нейтральную и комфортную среду для членов общества. «Потенциальная агрессия»² социума уравнивается принятыми правилами речевого этикета и, в частности, — вежливостью в речи. Развитие возможностей и каналов для коммуникации, появление всё более усовершенствованных систем искусственных голосовых ассистентов заставляет обратиться к проблеме просодических механизмов, организующих вежливое высказывание.

Пилотный эксперимент

В декабре 2023 года был проведен пилотный эксперимент. Его цель заключалась в выявлении первичных акустических ключей вежливости/невежливости.

В опросе приняли участие 22 носителя русского языка в возрасте 18–30 лет (82%), 41–51 год (4%), старше 51 года (14%). Большинство из них женщины (95%). 18 человек (81%) имеют филологическое образование, остальные специализируются в биологии, международной журналистике и международных отношениях.

Структура эксперимента

Опрос был разделен на две части. Предварительно респонденты указали возраст, пол, наличие/отсутствие филологического образования.

В первой части респондентам предлагалось прослушать 3 аудиозаписи и отметить, насколько:

¹ Научный руководитель — канд. филол. наук, доцент, У.Е. Кочеткова.

² «Проблемой для любой социальной группы является контроль ее внутренней агрессии, вежливость предполагает потенциальное наличие агрессии, потому что вежливость стремится обезоружить ее» [3].

«Одна из социальных функций общения — подавление агрессии» [2].

1. Вежливо звучало высказывание.
2. Эмпатично звучало высказывание.
3. Хотелось бы продолжать общение с говорящим.

Предлагаемые аудиозаписи — отрывки из интервью 1) телеведущей и композитора, 2) разговор психолога с участником интернет-шоу и 3) политического деятеля на телевидении. В каждом монологе присутствовал маркер вежливости (*пожалуйста, спасибо*).

Во второй части респонденты прослушали 7 одинаковых фраз, записанных одним диктором, при этом перед диктором стояла установка произнести фразу:

1. Вежливо, с маленькой социальной дистанцией (обращение к другу).
2. Вежливо, с большой социальной дистанцией (обращение к профессору).
3. С просьбой, жалобно, умоляюще.
4. Раздраженно.
5. Нейтрально.
6. Невежливо, гневно.
7. Печально, уныло.

Сама фраза представляла собой лексически нейтральное предложение, без маркеров вежливости: *Остановка — прямо за углом*.

Носителям русского языка нужно было прослушать каждую запись и ответить на три вопроса:

- 1) насколько вежливой и эмпатичной они считают фразу;
- 2) насколько велика социальная дистанция между говорящим и его собеседником;
- 3) какие эмоции у них вызвала бы фраза, если бы она была обращена к ним.

Результаты пилотного эксперимента

Для доклада были отобраны и проанализированы наиболее яркие контексты, отмеченные респондентами как практически однозначно вежливые/невежливые и эмпатичные/неэмпатичные.

1 часть

В докладе будет рассмотрен первый контекст, который большинство респондентов описали как вежливый, неэмпатичный и с большой социальной дистанцией.

Кроме того, респонденты могли оставлять комментарии после записи. Особенно интересным является комментарий:

- «Неприятный акцент на «что» в словах «что это за премия», будто вопрос с претензией».

Обращаясь к рисунку движения ОТ, мы видим, что после указанного слога происходит резкое повышение тона. Данное слово реализовано с дополнительной просодической выделенностью (ДПВ) (которая обозначена в орфографической транскрипции как [+]). По системе интонационных конструкций Нины Борисовны Вольской [1] это высказывание относится к модели 08: *Пожалуйста, расскажите мне, [+]что это за [08]премия.*

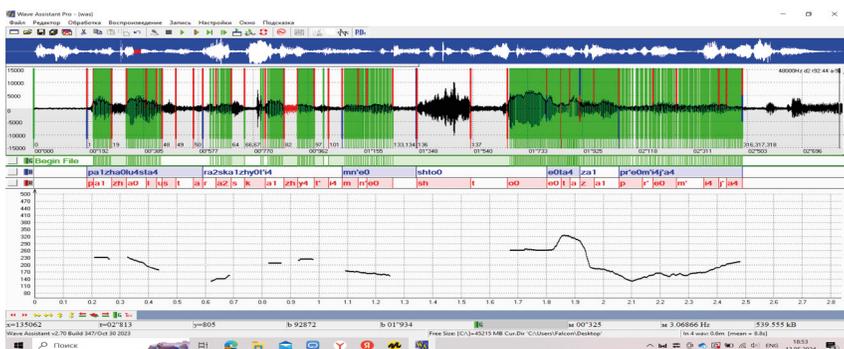


Рис. 1. График ОТ для фразы Пожалуйста, расскажите мне, [+]что это за [08] премия

Среднее значение ЧОТ в слоге [e0ta4], следующим за [shto0], значительно выше, чем в остальной части высказывания и равно 320 Гц (ср. после слога с ДПВ тон опускается на 14 полутонов, до 140 Гц. И делает резкий «скачок» на 3,5 полутона за 0,5 сек).

Таким образом можно заключить, что регистр высказывания может переходить в невежливый, из-за резких мелодических перепадов на гласных при ДПВ, а также при значительном повышении тона.

Далее будут рассмотрены результаты второй части опроса и выделены новые акустические характеристики вежливости/невежливости.

2 часть

Среди предложенных семи контекстов респонденты однозначно выделили две записи:

1. Контекст № 2 как вежливый, с большой социальной дистанцией между собеседниками и эмпатичный.

2. Контекст № 3 как невежливый и неэмпатичный, с трудно определяемой социальной дистанцией.

В остальных записях мнения респондентов существенно разошлись, поэтому для анализа выбраны записи № 2 и № 3 из второй части опроса.

Акустический анализ проведен вручную при помощи программы Praat и Wave Assistant.

Длительность

Фраза *Остановка* — прямо за углом насчитывает 9 гласных и 12 согласных фонем. В таблицах 2 и 3 представлена длительность слогов и длительность отдельных фонем в двух вариантах реализации одной и той же фразы. В таблицах 1 и 2 представлена длительность слогов и длительность отдельных фонем в двух вариантах реализации одной и той же фразы.

Таблица 1. Распределение длительности слогов в контексте № 2 (вежливый, с большой социальной дистанцией, эмпатичный)

	Слог		Длительность (сек)	Гласные	Длительность (сек)	Согласные	Длительность (сек)
	[a2]	V	0.073	[a2]	0.073	[s]	0.075
	[sta1]	CCV	0.228	[a1]	0.067	[t]	0.087
	[no0]	CV	0.124	[o0] [+]	0.063	[n]	0.063
	[fka4]	CCV	0.147	[a4]	0.026	[f]	0.067
	[pr'a0]	CCV	0.302	[a0] [+]	0.117	[k]	0.053
	[ma4]	CV	0.136	[a4]	0.063	[p]	0.128
	[za0]	CV	0.154	[a0]	0.074	[r']	0.057
	[u1]	V	0.123	[u1]	0.123	[m]	0.071
	[glo0m]	CCVC	0.291	[o0] [ИЦ]	0.074	[z]	0.081
						[g]	0.04
						[l]	0.083
						[m]	0.097
Средняя длительность			0.175		0.076		0.075
Общая длительность фразы							1.599

Таблица 2. Распределение длительности слогов в контексте № 3
(невежливый, неэмпатичный, с трудно определимой социальной дистанцией)

	Слог		Длительность (сек)	Гласные	Длительность (сек)	Согласные	Длительность (сек)
	[a2]	V	0.045	[a2]	0.045	[s]	0.056
	[sta1]	CCV	0.193	[a1]	0.058	[t]	0.079
	[no0]	CV	0.171	[o0] [+]	0.098	[n]	0.073
	[fka4]	CCV	0.172	[a4]	0.035	[f]	0.048
	[pr'a0]	CCV	0.276	[a0] [+]	0.14	[k]	0.089
	[ma4]	CV	0.14	[a4]	0.054	[p]	0.069
	[za0]	CV	0.161	[a0]	0.076	[r']	0.051
	[u1]	V	0.166	[u1]	0.166	[m]	0.086
	[glo0m]	CCVC	0.518	[o0] [ИЦ]	0.144	[z]	0.086
						[g]	0.028
						[l]	0.17
						[m]	0.176
Средняя длительность			0.205		0.091		0.084
Общая длительность фразы							1.854

При сравнении двух вариантов реализации фразы можно заметить, что (1) в невежливом неэмпатичном варианте общая длительность фразы, а также составляющих ее слогов и отдельных звуков в среднем выше, чем в вежливой эмпатичной реализации; (2) в обоих вариантах отсутствуют паузы; (3) в вежливом эмпатичном варианте интонационный центр смещается на предпоследний слог (длительность слога больше, чем в контексте № 3), в то время как в контексте № 3 интонационным центром является последний слог.

Интонационная модель, интонационный центр и ЧОТ

Обе фразы характеризуются тремя словами с дополнительной просодической выделенностью ([a2sta1no0fka4], [pr'a0ma4], [u1glo0m]). Интонационным центром является последнее слово фразы [u1glo0m].

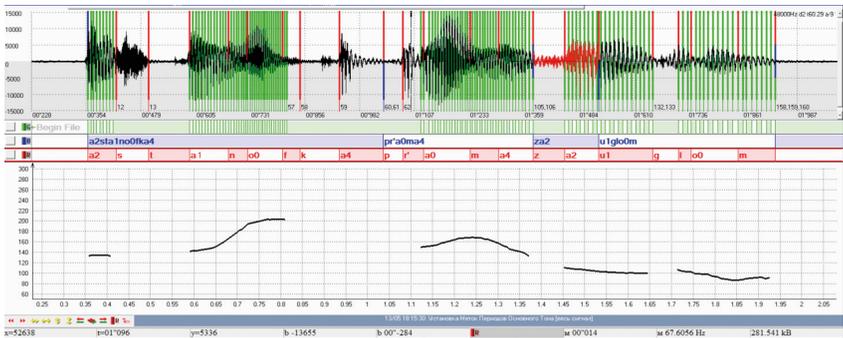


Рис. 2. График ОТ для контекста № 2 (вежливый, с большой социальной дистанцией, эмпатичный)

Согласно системе интонационных моделей Н. Б. Вольской в данном варианте прочтения фразы *Остановка — прямо за углом* реализована модель 01a (завершенность, соответствует ИК1 Е. А. Брызгуновой).



Рис. 3. График ОТ для контекста № 3 (невежливый, неэмпатичный, с трудно определимой социальной дистанцией)

По системе интонационных моделей Н. Б. Вольской контекст № 3 относится к модели 01b — неполная завершенность.

В таблице 3 описано движение ОТ в интонационном центре фразы и в слогах с дополнительной просодической выделенностью.

Также были составлены графики сравнения движения ОТ в этих слогах в двух контекстах.

Как можно видеть на графиках, рисунок ЧОТ различен для описываемых контекстов:

1. Слог [no] при произнесении в вежливой эмпатичной фразе имеет более низкое начало, более высокий конец и характери-

Таблица 3. Движение ЧОТ в интонационном центре фразы и в слогах с дополнительной просодической выделенностью.

Слог	Контекст	Вид слога	Движение тона
[po0]	№ 2 (вежливый, с большой социальной дистанцией, эмпатичный)	С дополнительной просодической выделенностью	Повышение от среднего к высокому регистру (≈ 2 полутона)
	№ 3 (невежливый, неэмпатичный, с трудно определяемой социальной дистанцией)	С дополнительной просодической выделенностью	Повышение от среднего к высокому регистру ($\approx 2,5$ полутона)
[pr'a0]	№ 2 (вежливый, с большой социальной дистанцией, эмпатичный)	С дополнительной просодической выделенностью	Повышение от среднего к высокому регистру (≈ 2 полутона)
	№ 3 (невежливый, неэмпатичный, с трудно определяемой социальной дистанцией)	С дополнительной просодической выделенностью	Повышение от среднего к высокому регистру (≈ 3 полутона)
[glo0m]	№ 2 (вежливый, с большой социальной дистанцией, эмпатичный)	Интонационный центр	Ровный низкий регистр (≈ 2 полутона)
	№ 3 (невежливый, неэмпатичный, с трудно определяемой социальной дистанцией)	Интонационный центр	Понижение от среднего до низкого регистра (≈ 8 полутонов)

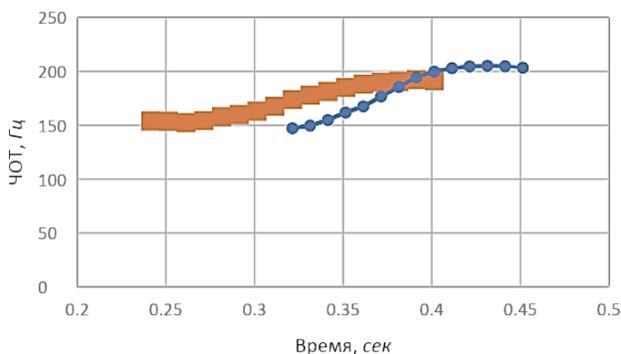


Рис. 4. Движение ОТ в слоге [po0], сравнение контекста № 2 (круги) и № 3 (квадраты)

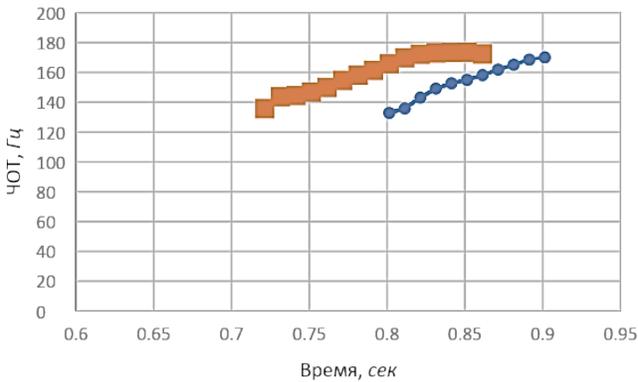


Рис. 5. Движение ОТ в слоге [pr'a0], сравнение контекстов № 2 (круги) и № 3 (квадраты)

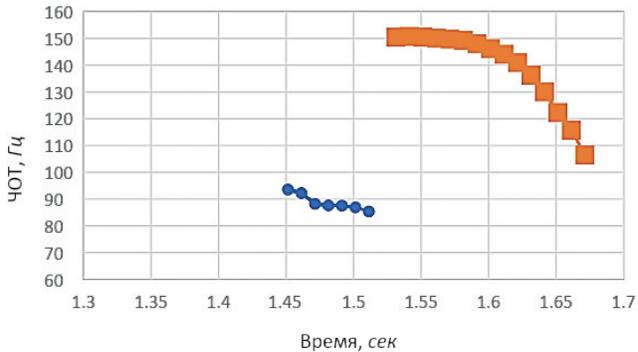


Рис. 6. Движение ОТ в слоге [glo0m], сравнение контекстов № 2 (круги) и № 3 (квадраты)

зуется более выразительным волнообразным движением тона из среднего регистра в высокий.

2. Слог [pr'a0] оформляется практически одинаковым рисунком ОТ.
3. Последний слог [glo0m] имеет самые яркие различия в движении ОТ. В невежливом, неэмпатичном высказывании тон падает от среднего регистра к низкому (от 150 Гц до 95 Гц, на 8 полутонов), при этом гласный удлинён. В вежливом эмпатичном высказывании тон характеризуется относительной стабильностью и низким регистром (от 95 Гц до 55 Гц, на 2 полутона).

Выводы

На основании полученных данных и проведенного анализа можно сделать следующие наблюдения:

1. Для вежливого контекста с большой социальной дистанцией характерны более быстрый темп речи и более низкий регистр произнесения, отсутствие пауз.
2. Для невежливого характерны резкие мелодические перепады, повышение мелодического регистра речи, растягивание гласных в просодически выделенных словах или ИЦ.

Таким образом, говорящий стремится сделать речь более плавной и непрерывной в вежливых высказываниях. Возможно, тенденция к ускорению темпа обусловлена желанием быть «удобным» собеседнику — не тратить его время и быстрее продемонстрировать знание норм социального общения.

Список использованной литературы

1. *Вольская Н. Б., Скрелин П. А.* Система интонационных моделей для автоматической интерпретации интонационного оформления высказывания: функциональные и перцептивные характеристики // Третий междисциплинарный семинар «Анализ разговорной русской речи» АРЗ — 2009. СПб., 2009. С. 28–40.
2. *Стернин И. А.* Введение в речевое воздействие. Воронеж, 2001. 227 с.
3. *Brown P., Levinson S.* Politeness: Some Universals in Language Usage. Cambridge: Cambridge University Press, 1987. 360 p.

ФОНЕТИЧЕСКИЕ ХАРАКТЕРИСТИКИ РЕЧИ ПРИ КОГНИТИВНОЙ НАГРУЗКЕ¹

Максимова Мария Романовна

(Санкт-Петербургский государственный университет)

Введение

Когнитивная нагрузка — это нагрузка на когнитивную систему человека при выполнении какой-либо задачи. Когнитивная нагрузка отражает объем ресурсов, затрачиваемых рабочей памятью человека. Рабочая память обеспечивает хранение и обработку информации. Объем и возможности рабочей памяти человека ограничены. Следовательно, для эффективного выполнения задачи необходимо контролировать нагрузку на рабочую память [10].

Существуют разные способы выявления когнитивной нагрузки: отслеживание взгляда (айтрекинг), электроэнцефалография, электрокардиография, измерение частоты сердечных сокращений, анализ характеристик голоса и речи [1].

Распознавание когнитивной нагрузки в речи применяется во многих сферах общественной жизни, в том числе в сфере образования. Выявление объема нагрузки на когнитивную систему обучающихся и соответствующая адаптация образовательного процесса позволяют повысить эффективность образовательных методик [10].

В современности распознавание когнитивной нагрузки позволяет повысить удобство использования диалоговых систем. Это особенно важно для работы голосовых ассистентов в ситуациях повышенной опасности. При распознавании признаков умственной нагрузки в речи пользователя система сможет адаптироваться под его психическое и физиологическое состояние [4].

При когнитивной нагрузке увеличивается напряжение голосовых складок и речевого тракта. Кроме того, наблюдается учащение дыхания. В результате, повышается подвязочное давление. Это приводит к повышению частоты основного тона (ЧОТ) [6]. Другими маркерами когнитивной нагрузки в речи могут быть повышение интенсивности [5; 6], сужение диапазона ЧОТ [4; 5; 6], ларингализация [4; 6; 11], динамика значений формант [6; 10; 11], изменение темпа речи и скорости артикуляции [4; 6; 7; 11].

¹ Научный руководитель — доцент СПбГУ, канд. филол. наук В. В. Евдокимова.

В исследованиях отмечаются как повышение, так и понижение определенных формант. Темп речи и скорость артикуляции могут повышаться или понижаться в зависимости от задач, выполняемых информантами, и от условий работы над заданиями. Например, исследование К. Мюллера [7] показало, что при ограничении по времени скорость артикуляции повышается. Однако выполнение дополнительной задачи приводит к понижению скорости артикуляции. В эксперименте Т. Ф. Япа [11] было выявлено, что темп речи повышается при высокой когнитивной нагрузке. Тем не менее, при низком и среднем уровне когнитивной нагрузки не наблюдается изменений темпа речи. Участники перцептивного эксперимента, проведенного Ф. Н. Ли [6], как правило, определяли средний уровень когнитивной нагрузки по медленному, но неравномерному темпу речи. Показателем высокого уровня когнитивной нагрузки был быстрый, но равномерный темп.

Целью данной работы является сравнение фонетических характеристик речи при наличии когнитивной нагрузки, при ее отсутствии и при предпаузальном удлинении.

Под предпаузальным удлинением понимается увеличение длительности звуков, расположенных рядом с границей синтагмы или фразы [2].

Материал

Материалом для исследования послужили студийные аудиозаписи речи семи студентов кафедры фонетики и методики преподавания иностранных языков СПбГУ. В качестве задачи, вызывающей когнитивную нагрузку, информантам было предъявлено 20 названий цветов, где цвет шрифта отличался от значения слова. Испытуемые должны были назвать цвет шрифта.

Этот эксперимент был основан на тесте Струпа (тесте словесно-цветовой интерференции) [9]. Тест Струпа — это нейропсихологический тест, с помощью которого оценивают способность человека подавлять реакцию на один из стимулов, если это затрудняет обработку второго стимула. Оба стимула подаются испытуемому одновременно. В данном случае речь идет о подавлении вербальных функций, ответственных за чтение слова и понимание его значения. Это необходимо, так как вербальные функции затрудняют выполнение сенсорно-перцептивных функций, обеспечивающих распознавание цвета шрифта [8; 9]. Дополнительной задачей были вопросы, связанные с теоретической фонетикой английского языка.

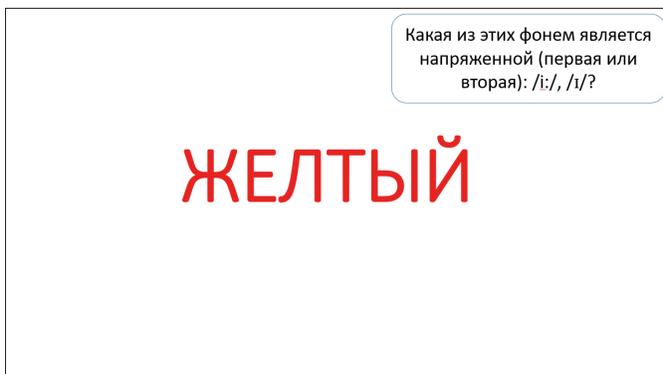


Рис. 1. Пример слайда с заданиями

Речь в случае отсутствия когнитивной нагрузки была реализована информантами при чтении фонетически представительного текста «Был тихий серый вечер» [3]. Общая длительность материала составила 49 мин 42 с.

Методика

Был проведен акустический анализ полученных записей с помощью программы Praat. Значения ЧОТ и длительностей были вычислены с помощью скриптов из коллекции SpeCT (The Speech Corpus Toolkit for Praat). Было проведено сравнение длительностей безударных гласных, щелевых согласных, сонантов и ларингализованных участков при когнитивной нагрузке, при отсутствии размышлений и при предпаузальном удлинении. Речь при когнитивной нагрузке была обнаружена только при выполнении информантами заданий. Речь при отсутствии затруднений наблюдалась и при прохождении теста, и при чтении текста. Предпаузальное удлинение рассматривалось только в чтении. Сравнивались нормализованные значения длительности с учетом коэффициента темпа речи, вычисленные по следующей формуле [2]:

$$\underline{d}_i = \frac{d_i - \alpha \mu_p}{\alpha \sigma_p}$$

где \underline{d}_i — нормализованная длительность звука, d_i — измеренная длительность звука, μ_p и σ_p — среднее значение и стандартное отклонение длительности данного типа звука по всему материалу для данного диктора, α — коэффициент темпа речи. Для вычисления коэффициента темпа речи была использована формула [2]:

$$\alpha = \frac{1}{N} \sum_{i=1}^N \frac{d_i}{\mu_{p_i}}$$

где N — количество звуков в рассматриваемом фрагменте, d_i — длительность i -того звука во фрагменте, μ_{p_i} — средняя длительность данного типа звука.

Также были сравнены диапазоны ЧОТ в интонационных контурах при перечисленных условиях. Значимость различий была выявлена с помощью двухвыборочного t -теста с различными дисперсиями.

Результаты и выводы

Результаты показали, что длительность безударных гласных в случае затруднения при выполнении задачи превышает длительность безударных гласных при отсутствии размышлений и при предпаузальном удлинении. Однако различия в длительности гласных при перечисленных условиях не были значимыми. Длительность щелевых согласных при когнитивной нагрузке также была выше, чем при отсутствии затруднений и в случае предпаузального удлинения. В длительности сонантов наблюдалась такая же закономерность. Различия в длительности щелевых согласных и сонантов при когнитивной нагрузке по сравнению с их длительностями в случае отсутствия затруднений и при предпаузальном удлинении чаще всего были значимы. Значимость различий была выявлена с помощью двухвыборочного t -теста с различными дисперсиями.

Чаще всего в случае затруднений при выполнении задания дикторы реализовали восходящее движение ЧОТ. Также частотным был восходяще-нисходящий интонационный контур. Диапазон значений ЧОТ в восходящих контурах при когнитивной нагрузке был значимо шире, чем диапазон восходящего мелодического движения при чтении неколических синтагм в фонетически представительном тексте.

Длительность ларингализованных участков при когнитивной нагрузке была значимо больше, чем в случае предпаузального удлинения.

Заключение

Увеличение длительности гласных и согласных при когнитивной нагрузке позволяет сделать вывод, что скорость артикуляции, в целом, понижается при данном условии. Увеличение диапазона ЧОТ в случае затруднений при выполнении задания противоречит результатам некоторых предшествующих работ [5; 6]. Следовательно, влияние ког-

нитивной нагрузки на диапазон ЧОТ нуждается в дополнительных исследованиях.

Список использованной литературы

1. *Еременко Ю. А., Залата О. А.* Психофизиологические подходы к проектированию образовательного контента в иммерсивной среде // Вопросы образования. 2020. С. 207–231.
2. *Качковская Т. В.* Взаимодействие сегментных и просодических факторов, влияющих на степень и локализацию предпаузального удлинения в русском языке. Автореф. дис. ... канд. филол. наук. СПб., 2015. 22 с.
3. *Степанова С. Б.* Фонетические свойства русской речи: реализация и транскрипция. Дис. ... канд. филол. наук. Л., 1988.
4. *Berthold A., Jameson A.* Interpreting Symptoms of Cognitive Load in Speech Input // UM99, User modeling: Proceedings of the seventh international conference. Springer, Vienna, 1999. P. 235–244.
5. *Huttunen K., Keränen H., Väyrynen E., Pääkkönen R., Leino T.* Effect of cognitive load on speech prosody in aviation: Evidence from military simulator flights // Applied ergonomics. 2011. P. 348–357.
6. *Le P. N.* The Use of Spectral Information in the Development of Novel Techniques for Speech-Based Cognitive Load Classification/ PhD thesis. Sydney, 2012.
7. *Müller C., Großmann-Hutter B., Jameson A., Rummer R., Wittig, F.* Recognizing time pressure and cognitive load on the basis of speech: An experimental study // International Conference on User Modeling. Berlin, Heidelberg, 2001. P. 24–33.
8. *Scarpina F., Tagini S.* The stroop color and word test // Frontiers in Psychology. 2017. <https://doi.org/10.3389/fpsyg.2017.00557>
9. *Stroop J. R.* Studies of interference in serial verbal reactions // Journal of Experimental Psychology. Vol. 18 (6). 1953. P. 643–662.
10. *Sweller J.* Cognitive load during problem solving: Effects on learning // Cognitive Science. Vol. 12 (2). 1988. P. 257–285.
11. *Yap T. F.* Speech Production Under Cognitive Load: Effects and Classification. PhD thesis. Sydney, 2012.

КОРЕЙСКО-РУССКАЯ ФОНЕТИЧЕСКАЯ ИНТЕРФЕРЕНЦИЯ¹

Пахомова Марина Алексеевна

(Санкт-Петербургский государственный университет)

Введение

Мысль о том, что мы воспринимаем иностранный язык через призму родного, не является новой. Наложение одной языковой системы на другую принято называть межъязыковой интерференцией. Это явление наблюдается на всех уровнях языка, в том числе и на фонетическом. Описывая то, что мы сейчас называем фонетической интерференцией, Е. Д. Поливанов писал: «Слыша незнакомое иностранное слово <...> мы пытаемся обнаружить в нем комбинацию наших фонологических представлений, разложить его на фонемы нашего родного языка и даже сделать это в согласии с законами следования фонем, присущими нашему языку» [3].

В данной статье приведено исследование корейско-русской фонетической интерференции на сегментном уровне. Несмотря на растущую популярность корейского языка в последние годы, работ, посвященных корейско-русской интерференции, немного. Кроме того, не все работы охватывают как консонантные, так и вокалические проявления интерференции [2]. Задачей данного исследования было рассмотреть интерференцию на сегментном уровне в целом, не ограничиваясь лишь некоторыми фонемами.

Основной целью работы стало определение существенных различий на сегментном уровне фонетических систем русского и корейского языков, а также выявление сегментных единиц, при реализации которых наиболее ярко проявляется интерференция.

Для достижения поставленной цели были решены следующие задачи:

- 1) обзор научной литературы, посвященной фонетической интерференции и описывающей фонетические системы русского и корейского языков;
- 2) сравнение двух фонетических систем на основе их описания, а также прогноз наиболее вероятных проявлений интерферен-

¹ Научный руководитель — доцент СПбГУ, канд. филол. наук С. О. Тананайко.

ции — ошибок, которые могут совершить в речи корейцы, изучающие русский язык;

- 3) запись речи дикторов-носителей корейского (спонтанная речь и чтение фонетически представительного текста);
- 4) проведение слухового, акустического, статистического анализа полученного материала;
- 5) общие выводы о результатах исследования.

Практическая значимость работы заключается в том, что результаты исследования можно будет использовать для разработки методики преподавания фонетики русского языка корейцам.

Фонетическая система корейского языка

Для того, чтобы природа интерференции была более очевидна, представляется необходимым сжато описать главные различия фонетических систем корейского¹ и русского языка.

Консонантизм

В отличие от русского языка, в корейской фонологии отсутствует различие согласных по: 1) твердости–мягкости; 2) глухости–звонкости.

По этой причине прогнозировались ошибки, связанные с твердостью–мягкостью и глухостью–звонкостью. А именно: замена мягких согласных на твердые или недостаточно мягкие, а также полное или частичное оглушение звонких согласных.

Однако в корейском языке есть противопоставления согласных, которых, наоборот, нет в системе русского языка: 1) по степени придыхания; 2) по напряженности–ненапряженности. Под напряженностью понимается напряжение мускулатуры гортани: сужение голосовой щели, не достигающее, однако, до полной смычки [6].

Корейский боковой аппроксимант /l/ в виде своего основного аллофона реализуется альвеолярно или ретрофлексно [5], а также обладает одноударным аллофоном [ɭ]. Ожидалось, что корейцы могут путать /l/ и /ɭ/ (и их мягкие пары), реализовывать /l/ альвеолярно (или ретрофлексно) и недостаточно веляризованно, а /ɭ/ заменять на одноударный [ɭ].

¹ В данной работе рассматривается литературная норма корейского языка, основанная на сеульском диалекте.

Таблица 1. Система согласных корейского языка¹

	Губно- губные	Переднеязыч- ные	Альвеолярно- палатальные	Велярные	Гортанные
Смычные взрывные	p, p*, p ^h	t, t*, t ^h		k, k*, k ^h	
Щелевые		s, s*			h
Аффрикаты			tc, tc*, tc ^h		
Носовые сонанты	m	n		ŋ	
Плавные		l			

Консонантная система корейского языка не отличается разнообразием щелевых согласных: из 19 согласных фонем только 3 являются щелевыми — /s/, /s*/, /h/. Стоит также упомянуть аллофон /s/, который встречается в позиции перед /i/ и йотированными, — альвеолярно-палатальный щелевой [ɕ]. В связи с этим ожидалось, что все щелевые русского языка, кроме /s/, будут попадать в зону интерференции. В частности, заднеязычный /x/ будет заменяться на гортанный /h/, а /z, ž/ будут подвергаться аффрикатизации и, вероятно, заменяться на аффрикату /tc/.

Вокализм

В отличие от русского языка, в корейском есть как монофтонги, так и дифтонги. Монофтонгов всего семь: /a/, /o/, /ʌ/, /ɛ/, /u/, /ʉ/, /i/.

Корейская фонема /ʉ/, в отличие от похожей на нее русской /i/, не является дифтонгоидной и реализуется с чуть более задним положением языка.

Русской гласной /o/ могут быть сопоставлены две гласные корейского языка:

- /o/ — среднего подъема, заднего ряда, огубленная. Отличается от соответствующей русской /o/ гораздо более сильной огубленностью и закрытостью;
- /ʌ/ — среднего подъема, заднего ряда, неогубленная. В отличие от русской /o/, она более открытая и неогубленная.

¹ Диакритическим знаком / * / обозначается напряженность согласного, а знаком / ^h / — сильная степень придыхания.

Материал и методика

Данное исследование проводилось на основе записей русской речи пяти носителей корейского языка: двоих мужчин и трех женщин. Возраст дикторов варьировался от 19 до 25 лет.

В ходе исследования использовался материал двух типов: чтение фонетически представительного текста и спонтанная речь.

Прежде всего дикторам было предложено прочитать фонетически представительный текст «Был тихий серый вечер» (далее — «ФПТ» или «текст»). Этот текст был создан на кафедре фонетики и методики преподавания иностранных языков филологического факультета СПбГУ. Он отражает фонетическую систему русского языка в целом [1].

Всего было получено 5 аудиозаписей с чтением ФПТ и 5 аудиозаписей со спонтанной речью — диалогом. Общая длительность полученного аудиоматериала составила примерно 70 минут.

Далее был отобран материал для анализа: анализировались только разборчиво произнесенные слова. Ошибками считались нарушения орфоэпической нормы, предположительно вызванные интерференцией. Таким образом, не были учтены ошибки, не являющиеся следствием интерференции. Например, не учитывались ошибки связанные с неправильной постановкой ударения, реализацией Ё вместо Е и наоборот.

Отобранный материал был сегментирован на фразы в программе анализа и обработки звукового сигнала Praat. Было создано два уровня аннотации: уровень с идеальной транскрипцией и уровень с реальной транскрипцией.

Для статистического анализа ошибок была написана программа на языке Python: программа считывала оба уровня аннотации, сравнивала идеальную и реальную транскрипции и выводила файл Excel с таблицей, содержащей нормативную фонему (из идеальной транскрипции), ошибочную реализацию этой фонемы (из реальной транскрипции), количество и процент каждой ошибочной реализации. Последующие статистические расчеты проводились с использованием программы Excel.

Кроме того, был проведен акустический анализ попавших в зону интерференции гласных фонем /o/, /u/, /i/.

Методом случайной выборки из речи каждого из пяти дикторов было отобрано по пять слов, содержащих интересующий нас гласный звук в ударной позиции между твердыми согласными. Для анализа гласных был создан отдельный уровень аннотации в программе Praat, а также написан скрипт, считающий средние значения первой и второй формант (далее — «F1» и «F2» соответственно) на интервалах с гласными. Таблица, которую выводил скрипт, была перенесена в программу Excel, где были проведены дальнейшие статистические расчеты: были

получены средние значения формант гласных /o/, /u/, /i/ отдельно по каждому диктору, отдельно по дикторам мужского и женского пола, по всем пяти дикторам. Также были выведены минимальные и максимальные средние значения по всем дикторам.

Для большей объективности эксперимента формантные значения /o/, /u/, /i/ в реализации корейских дикторов были сопоставлены с формантами этих же гласных в реализации русских дикторов. Для этого использовались записи чтения фонетически представительного текста «Был тихий серый вечер» пятью дикторами, являющимися носителями русского языка: два диктора были мужского пола и три диктора — женского пола. Было отобрано по 10 реализаций гласных /o/, /u/, /i/ (в позиции под ударением между твердыми согласными) у дикторов мужского пола и столько же у дикторов женского пола. На этом материале были посчитаны средние значения формант /o/, /u/, /i/ в речи русских дикторов аналогично тому, как была получена статистика частотных характеристик этих гласных в речи дикторов-корейцев.

Частотные характеристики (F1, F2) фиксировались на стационарном участке гласного так, чтобы влияние процессов коартикуляции на среднее значение формант было минимальным. Ошибки в виде неверных значений формант в программе Praat выявлялись и исключались из статистики.

Результаты и выводы

Таблица 2. Фонемы русского языка, при реализации которых корейцы совершали больше всего ошибок

	Русская фонема	Процент неверной реализации (ФПТ)		Русская фонема	Процент неверной реализации (спонтанная речь)
1	ž	90,9	1	ž	73,9
2	d'	80,0	2	d'	69,2
3	š	62,3	3	z'	60,0
4	z	62,0	4	g	51,1
5	g'	60,0	5	b	50,0
6	b	51,9	6	š	47,4
7	g	51,8	7	z	43,8
8	z'	50,0	8	x	43,2
9	d	47,9	9	t'	41,7
10	b'	47,6	10	l'	41,5

Результаты слухового, акустического и статистического анализа согласных звуков показали, что наиболее частыми типами ошибок в реализации согласных стали следующие явления:

- Оглушение звонких согласных;
- Недостаточно мягкая реализация мягких согласных;
- Смягчение /š, ž, c/;
- Замена /č/ на /tɕ/, а /šː/ — на [ɕ];
- Аффрикатизация /ž, z, z', t', d'/;
- Альвеолярная или ретрофлексная реализация /l/;
- Щелевая артикуляция /p, b/ и смычная артикуляция /f, v/;
- Щелевая артикуляция /g/ в спонтанной речи.

Большая часть представленных ошибок соответствует прогнозу, который был сделан после сопоставления фонетических систем корейского и русского языков. Эти ошибки, очевидно, являются следствием корейско-русской интерференции. Тем не менее, встречались и непредвиденные ошибки: щелевая артикуляция русских согласных /f, v, g/. Остановимся на них подробнее.

Поскольку в корейском языке нет щелевых губно-зубных (равно как и щелевых губно-губных) фонем, ожидалось, что корейцы будут заменять /f, v/ на губно-губные смычные /p, b/. Это действительно имело место, однако при анализе материала встречалось и обратное явление: щелевая (губно-губная или губно-зубная) артикуляция на месте губных смычных /p/, /b/. Причиной этого является смешение двух типов губных: смычных и щелевых. Корейцы знают, что в русском языке есть смычные губные, но в речи не всегда реализуют русское фонемное противопоставление. В результате, корейцы «путают» все четыре фонемы: /f/, /v/ и /p/, /b/.

В спонтанной речи также была отмечена щелевая реализация заднеязычного на месте смычного: /g/ заменялся на заднеязычные или гортанные щелевые: [ɣ], [x], [ɦ], [h]. Так как как щелевая реализация /g/ наблюдалась только в спонтанной речи, можно связать такую артикуляцию со спецификой спонтанной речи и быстрым темпом речи. Согласно результатам исследования спонтанной речи на материале русского языка, описанным в «Фонетике спонтанной речи» [4], спирализация (ослабление смычки) является «самой частой модификацией смычных согласных». С учетом этого звонкую щелевую артикуляцию /g/ в спонтанной речи носителей корейского языка не следует трактовать только как следствие интерференции. Глухая щелевая реализация (например, [x], [h] вместо соответствующих звонких), однако, будет являться результатом интерференции.

В том, что касается неверной реализации гласных, слуховой анализ материала показал, что из всех гласных русского языка, чаще всего в зону интерференции попадали только три: /o/, /u/, /i/. Для подтверждения этого был проведен акустический анализ интересных нас гласных.

1. Акустический анализ гласного /u/.

По признаку подъема гласный /u/ в реализации дикторов-корейцев в целом можно считать соответствующим норме: средние показатели по всем дикторам отличаются несущественно — на 17 Гц (у русских дикторов — 413 Гц, у корейских дикторов — 396 Гц). Однако средние значения по дикторам мужского пола (426 Гц и 366 Гц у русских и корейцев соответственно) показывают, что корейцы-мужчины реализовывали более закрытые /u/.

По ряду /u/ в реализации дикторов-корейцев более передний, чем у носителей русского языка: средние значения различаются более чем на 100 Гц. У носителей русского языка — 1066 Гц, у носителей корейского — 1202 Гц. Показательны и максимальные значения F2: среди русских дикторов максимальное значение составило 1274 Гц, а среди корейцев — 1541 Гц, что говорит о довольно значительном продвижении /u/ вперед.

2. Акустический анализ гласного /o/.

Как и ожидалось при прогнозе проявлений интерференции, /o/ иногда заменялся на более огубленный и закрытый гласный. Этим обусловлены несколько более низкое среднее значение F1: 500 Гц у русских дикторов и 462 Гц у носителей корейского. Минимальное значение также оказалось ниже: 401 Гц и 340 Гц у русских и корейцев соответственно.

Что касается признака ряда, имело место продвижение /o/ вперед, однако не такое значительное, как при реализации /u/: разница средних значений F2 не превышает 100 Гц. У русских дикторов среднее значение составило 957 Гц, у дикторов-корейцев — 1022 Гц.

3. Акустический анализ гласного /i/.

Интерференция на гласном /i/ проявлялась по-разному: имели место как замены /i/ на [i]-образный звук, так и замены /i/ на корейский монофтонг заднего ряда — /ɨ/.

Средние значения F1 (408 Гц и 437 Гц в речи русских и корейцев соответственно) различались несущественно, поэтому можно говорить о том, что по признаку подъема /i/ в реализации носителей корейского является нормативным. И все же стоит отметить, что в некоторых случаях /i/ был более открытым — максимальное значение F1 среди корей-

ских дикторов гораздо выше, чем среди русских: 470 Гц в реализации русских и 621 Гц в реализации корейцев.

Широкий диапазон значений F2 гласного /i/ объясняется дифтонгоидным характером этой фонемы. В речи русских дикторов F2 варьируется от 1179 Гц до 2321 Гц. В речи носителей корейского — от 1051 Гц до 2242 Гц. Более низкий порог значений F2 вызван влиянием корейской вокалической системы: а именно — заменой русской фонемы /i/ на корейскую /ш/. Тем не менее, как показал анализ реализаций /i/ в речи носителей русского языка, более задняя монофтонгоидная реализация /i/ также встречается в русском языке.

Среднее значение F2 оказалось выше в реализации /i/ корейскими дикторами: 1674 Гц и 1760 Гц в реализации русских дикторов и дикторов-корейцев соответственно. Это связано с тем, что в некоторых случаях корейцы либо довольно сильно продвигали /i/ вперед, либо же полностью заменяли /i/ на гласный типа [i].

Заключение

В заключение можно сказать, что была достигнута главная цель работы: было проведено сравнение фонетических систем русского и корейского языков, определены основные виды ошибок, вызванные интерференцией, и конкретные фонемы русского языка, представляющие наибольшую трудность для носителей корейского.

Наиболее частыми ошибками в реализации согласных стали, в частности, следующие явления: оглушение звонких согласных; твердая реализация мягких согласных; смягчение /š, ž, c/; замена /č/ на /te/, а /š:/ — на [ɕ]; аффрикатизация /ž, z, z', t', d'/; альвеолярная или ретрофлексная реализация /л/ и др.

Среди гласных интерференции чаще всего подвергались /u/, /o/, /i/:

- 1) /u/ в реализации носителей корейского языка был более передним, чем нормативный гласный;
- 2) Русский /o/ иногда заменялся на гораздо более огубленную и закрытую корейскую фонему /o/;
- 3) /i/ сближался с /i/ или полностью заменялся на него, либо же, наоборот, отодвигался назад, сближаясь с корейским монофтонгом заднего ряда /ш/.

Список использованной литературы

1. *Бондарко Л. В., Степанова С. Б.* Текстовый модуль «Фонетически представительный текст» // Бюллетень Фонетического фонда русского языка. № 14, сентябрь 1992. Vochum; Л., 1992.
2. *Ким Ё.* Особенности фонетической реализации согласных в русской речи корейцев. Дис. канд. филол. наук. СПб., 1998.
3. *Поливанов Е. Д.* Субъективный характер восприятий звуков языка // Статьи по общему языкознанию. М., 1968. С. 236–254.
4. Фонетика спонтанной речи / под. ред. Н. Д. Светозаровой. Л.: Изд-во ЛГУ, 1988.
5. *Crosby D., Dalola A.* “Phonetic Variation in the Korean Liquid Phoneme” // Proceedings of the Linguistic Society of America. Vol. 6 (1). 2021. P. 701–712.
6. *Shin J., Kiaer J., Cha J.* The Sounds of Korean. Cambridge: Cambridge University Press, 2012.

РУССКО-КИТАЙСКАЯ ИНТЕРФЕРЕНЦИЯ ПРИ РЕАЛИЗАЦИИ ИНТОНАЦИИ В РЕЧИ НА РУССКОМ ЯЗЫКЕ¹

*Пискунова Жанна Евгеньевна
(ФГАОУ «Санкт-Петербургский политехнический
университет Петра Великого»)*

Введение

В данной статье будет рассмотрена интерференция в русской речи китайских носителей, а именно перенесение контуров китайского языка в русскую речь, что влечет за собой нарушение нормы интонации русского языка.

Интерференция — взаимодействие систем различных языков в рамках двуязычия, которое складывается при языковых контактах или при самостоятельном овладении неродным языком; проявляется в отклонениях от принятой нормы и системы неродного языка под влиянием родного [7].

Необходимо также рассмотреть понятие «языковой интерференции» и «межъязыковой фонетической интерференции». Языковую интерференцию в узком значении можно определить как перенесение языковых характеристик родного языка в иноязычную коммуникацию [9]. Межъязыковая фонетическая интерференция отражает фонетический аспект рассматриваемого явления, в рамках данного понятия стоит обратить внимание на ритмические и интонационные модели в системах каждого из рассматриваемых языков [8].

Интонация складывается из сочетания особых акустических характеристик речи, таких как сила звука, движение тона, длительность и тембр голоса. Различное сочетание данных характеристик составляет смысловой и эмоциональный компоненты речи [3].

Материал

В русском языке интонация представлена несколькими контурами. Наиболее распространенными являются классификации Е. А. Брызгуновой [2] и Н. Б. Вольской [5]. В данной статье будет использована последняя классификация.

¹ Научный руководитель — доцент ФГАОУ «Санкт-Петербургский политехнический университет Петра Великого», канд. филол. наук М. П. Агафонова.

Н. Б. Вольская выделяет 13 интонационных моделей:

- Модель 1 — нисходящая интонация, оформляет конец абзаца, конец предложения, завершённые синтагмы;
- Модель 2 — логическое (эмфатическое) ударение;
- Модель 3 — синтагматическое ударение на вопросительном слове;
- Модель 4 — восклицания первого типа;
- Модель 5 — восклицания второго типа;
- Модель 6 — ровный тон (с многоточием);
- Модель 7 — общий вопрос (и вопросы с «разве»);
- Модель 8 — вопрос с «а» (восходящая интонация);
- Модель 9 — слова автора, пояснения, интонация вводности;
- Модель 10 — незавершённость при сочинительной связи, при двоеточии;
- Модель 11 — вопрос без вопросительного слова
- Модель 12 — незавершенность, выражение оценки в предложениях с местоименными словами;
- Модель 13 — незавершенность, установка на продолжение с эмоционально-модальными значениями типа сомнения, недоверие.

В китайском языке акцентное выделение реализуется при помощи иного порядка слов в высказываниях. Система ударения определяется тонами. В китайском языке выделяют 4 тона: ровный, нисходящий, восходящий и нисходяще-восходящий. Они и определяют интонационные контуры в речи. Частота тона в речи носителей постоянно меняется в процессе говорения, что не характерно для русскоговорящих. Монотонность, как и в русском языке, недопустима [10].

Среди наиболее частых ошибок в русской речи китайских носителей можно выделить следующие: трудности в определении и передаче интонационного центра в предложении, сложность в сохранении и передаче эмоциональности в речи, неверная сегментация, отклонения в акцентном выделении, а также ошибки в фразовом ударении, нарушение ритмического рисунка и темпа речи, избыточная паузация [1; 10].

Исходя из вышеперечисленных положений, было проведено исследование для подтверждения или опровержения выявленных ошибок.

Цель исследования: выявить наиболее частотные и устойчивые отклонения при реализации русской интонации носителями китайского языка.

Материал: записи чтения фонетически представительного текста «Был тихий серый вечер» [4].

Метод

К исследованию привлекли 12 дикторов, 4 из которых являлись студентами 1 курса Санкт-Петербургского Политехнического университета, направление Лингвистика. Срок изучения русского языка — 1 год. Уровень владения русским языком А1–А2 (порядковые номера дикторов для символьного обозначения: 1–4). Остальные 8 дикторов студенты 2 и 3 курса Санкт-Петербургского Политехнического университета, направление Лингвистика. Срок изучения русского языка 2–3 года. Уровень владения русским языком А2–В1 (порядковые номера дикторов для символьного обозначения: 5–12). Все респонденты являются студентами из Китая и проживают в России.

Методика: слуховой и акустический анализ в программе для фонетического анализа речи «Praat». Во время эксперимента было проанализировано 100 минут 12 секунд записей речи. Для каждой синтагмы было необходимо определить интонационную модель согласно классификации Н.Б.Вольской, особенности реализации мелодики предцентра и постцентра, паузации и деления на синтагмы, был подсчитан процент отклонений от интонационной нормы русского языка в речи каждого из дикторов.

Результаты и выводы

В ходе исследования и обработки полученных данных были получены следующие результаты.

Таблица 1. Процентное соотношение количества ошибочных реализаций

Ошибка (в процентах)/ диктор (номер)	1	2	3	4	5	6	7	8	9	10	11	12
Ошибки при выборе интонационных конструкций	70	76	64	67	54	7	45	94	94	95	0	8
Ошибки при выборе места интонационного центра	33	3	1	4	62	39	44	94	94	94	0	8
Ошибки при реализации предцентра/постцентра	100	82	82	58	22	15	13	90	22	31	0	100
Ошибки при делении на синтагмы	85	0	47	4	18	35	16	9	9	14	0	0
Ошибки паузации	100	36	48	15	23	24	17	9	9	0	0	0

Таблица 2. Процентное соотношение реализаций интонационных моделей

Интонационная модель (в процентах)/диктор (номер)	1	2	3	4	5	6	7	8	9	10	11	12
1a	0	1	3	5	0	24	32	7	7	9	3	73
1b	56	69	70	84	31	16	14	85	34	25	71	6
2	0	8	6	0	1	2	0	0	52	57	0	0
3	0	0	0	0	0	0	2	0	2	0	1	0
4	0	0	0	0	0	0	0	2	0	0	2	0
4b	0	0	0	0	0	0	2	0	0	0	0	0
7	0	2	0	0	6	0	5	0	0	0	0	0
7a	11	1	5	5	0	1	3	3	3	0	5	3
8	0	0	1	0	0	0	0	0	0	0	1	0
11	0	3	3	1	4	46	12	0	0	1	2	2
11a	33	16	12	5	56	3	28	3	2	8	15	16
12	0	0	0	0	1	0	0	0	0	0	0	0
13	0	0	0	0	0	1	2	0	0	0	0	0

Таким образом, исходя из полученных данных можно сделать следующие выводы. В основном, в речи дикторы реализуют не более 6–7 контуров, три из которых проявляются наиболее частотно. Речь не отличается разнообразием использования интонационных моделей. Дикторы используют контур 1b (неполная завершённость, практически ровный тон — установка на связность текста) чаще других. Также, с высокой частотностью используется контур 11a (восходящая интонация со сдвигом мелодического пика на следующий слог), реже используется контур 1a (полная завершённость — конец предложения). Дикторы 1–12 используют больше различных интонационных контуров, чем дикторы 1–4, но разница не сильно значительная.

Что касается типичных ошибок при реализации контуров, то наиболее частая ошибка при реализации интонационных моделей, это монотонность речи, отсутствие вопросительных интонаций, плохо сформированный навык интонирования. Дикторы 1–4 склонны к реализации в предцентре и постцентре только ровного тона, неверному делению речи на синтагмы. Также наблюдается чрезмерное наличие пауз хезитации (в основном, после каждого знаменательного слова). В речи дикторов 5–12 паузы хезитации встречаются реже, все же частотны

ошибки по месту реализации, отсутствию деklinации, диктор 10 склонен к чрезмерной выделенности.

Таблица 1 наглядно демонстрирует, что в речи дикторов 5–12 ошибки при реализации предцентра/постцентра, ошибки при делении на синтагмы, а также ошибки паузации встречаются гораздо реже. Однако, более часто встречаются ошибки при выборе места интонационного центра. Процентное соотношение по параметру ошибки при выборе интонационных конструкций у дикторов 1–4 и 5–12 примерно одинаково.

Однако, стоит отметить, что при чтении дикторы 1–12 верно определяют интонационный центр, а также выделяют акцентные единицы на знаменательных словах. Без переноса интонационного центра реализуются только те синтагмы, где интонационный центр приходится на последний слог, что является нормой в речи русскоговорящей молодежи. Дикторы 5–12 склонны к правильному делению речи на синтагмы, в речи диктора 6 выделяется высокая естественность речи.

Заключение

В данной статье было рассмотрено явление интерференции в русской речи носителей китайского языка с точки зрения интонационных особенностей. Интересно было рассмотреть частотность перенесения тонов из китайского языка в русский. Однако, как было выявлено в ходе исследования, в речи носителей китайского языка наоборот преобладает излишняя монотонность. Также были определены другие особенности ошибочных реализаций речи, такие как неверное определение предцентра и постцентра, чрезмерная паузация, неверное деление на синтагмы и др.

К исследованию было привлечено 12 дикторов, носителей китайского языка, обучающиеся в Санкт-Петербургском Политехническом университете. Их речь была записана и позже проанализирована в специальной программе «Praat». По результатам анализа можно сделать вывод, что в речи дикторов 5–12 заметны улучшения в интонационных способностях. Как в первой таблице, так и во второй по некоторым параметрам их показатели выше, чем у дикторов 1–4.

Тем не менее, некоторые ошибки, выявленные у дикторов 1–4, остаются в речи. Наиболее явно это отражено в Таблице 2, в которой показаны результаты анализа использования различных интонационных моделей носителями китайского языка в русской речи. Как можно заметить, дикторы 5–12 используют достаточно мало интонационных контуров в своей речи, ограничиваясь несколькими моделями. Некоторые ошибки в интонации также сохраняются в речи дикторов.

Подводя итоги полученным данным, стоит отметить, что данное исследование может лечь в основу материала для преподавателей РКИ. На выявленные типичные ошибки и редко используемые интонационные контуры следует обращать больше внимания в ходе обучения интонации русского языка и давать разнотипные упражнения на их отработку, а также уделять больше времени на практику и закрепление материала.

Список использованных источников

1. *Березовская Я. Л., Ду М.* Специфика мелодического контура высказываний в русском языке в восприятии носителей китайского языка: лингводидактический аспект // Вестник Челябинского государственного университета. 2020. № 7(441). С. 21–27.
2. *Брызгунова Е. А.* Звуки и интонация русской речи. М.: **Русский язык**, 1977.
3. *Брызгунова Е. А.* Практическая фонетика и интонация русского языка. М., 1963.
4. *Бондарко Л. В.* Текстовый модуль «Фонетически представительный текст» // Бюллетень Фонетического фонда русского языка. Vochum-Л. 1992. № 4. С. 132–134.
5. *Вольская Н. Б.* Система интонационных моделей для автоматической интерпретации интонационного оформления высказывания: функциональные и перцептивные характеристики // Анализ разговорной русской речи (АРЗ-2009): тр. третьего междисциплинарного семинара. СПб.: Изд-во С.-Петербург. гос. ун-та аэрокосмического приборостроения, 2009.
6. *Демидчик Л. В.* Фонетическая реализация и функциональная нагрузка восходяще-нисходящей интонации в речи русской молодежи (экспериментально-фонетическое исследование). СПб.: Изд-во С.-Петербург. ун-та, 2009.
7. Лингвистический энциклопедический словарь / под ред. В. Н. Ярцевой. М.: Сов. энцикл., 1990. 125 с.
8. *Любимова Н. А.* Фонетическая интерференция: учеб. пос. Л.: ЛГУ, 1985. 35 с.
9. *Тернин А. В.* Взгляд на языковую интерференцию и степени ее проявления // Современные проблемы науки и образования. 2014. № 3. URL: <http://www.science-education.ru/ru/article/view?id=13089> (дата обращения: 18.11.2021).
10. *Ши Д.* Анализ интонационных трудностей в русской речи китайцев // Филология и культура. 2020. № 4 (62). С. 76–82.

АУДИОВИЗУАЛЬНОЕ ОФОРМЛЕНИЕ ПРАГМАТИЧЕСКОГО МАРКЕРА В ОБЩЕМ¹

Радионова Евгения Алексеевна
(Санкт-Петербургский государственный университет)

Введение

На сегодняшний день мультимодальный анализ является одним из часто используемых методов изучения устной речи, поскольку перед исследователем стоит задача понять устройство дискурса во всей его полноте, то есть изучить не только содержание и просодическое оформление речи, но и то, что говорящий передаёт собеседнику через визуальный канал, а именно посредством использования различных жестов. При исследовании устной речи важно изучать её вербальную составляющую совместно с невербальной, так как, согласно Е. А. Гришиной, высказывание и сопровождающая его жестикауляция возникают в мозгу говорящего в момент формирования высказывания как единое целое и лишь на конечном этапе осуществления «расходятся» по двум разным модусам — по зрительному и собственно лингвистическому [4].

Как правило, жесты соотносятся со значением слова или высказывания, например, жест *пожать плечами* используется при вербализации выражения *не знаю* или указательный жест сопровождает использование местоименного наречия *туда, там (там посмотри)*.

В данном докладе в центре внимания оказывается анализ просодических характеристик ПМ в *общем* и жестикауляции, реализующейся при его произнесении, а не описание семантики этой единицы, так как прагматический маркер утратил лексическое значение, в результате чего выполняет в дискурсе лишь определенные функции [1].

Материал

Материалом для исследования служат 35 контекстов, в которых содержится ПМ в *общем*, взятых из мультимедийного подкорпуса национального корпуса русского языка (НКРЯ).

Для анализа были выбраны употребления ПМ в *общем* в функции разграничителя-гезитатива (термин Н. В. Богдановой-Бегларян [5]).

¹ Научный руководитель — доцент СПбГУ, канд. филол. наук У. Е. Кочеткова.

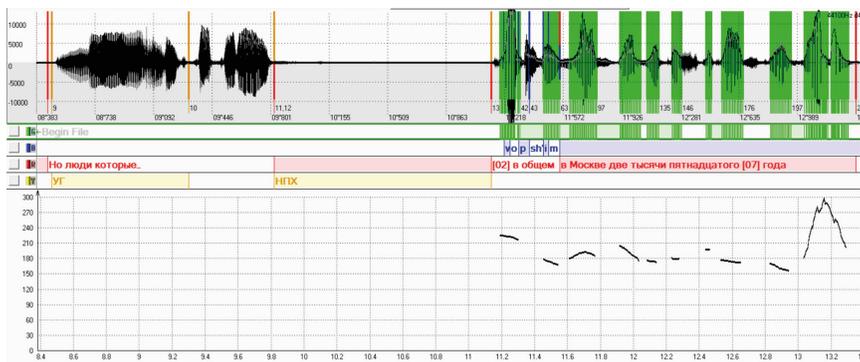


Рис. 1. Осциллограмма, расшифровка речи и интонационный контур контекста с маркером в *общем*

В рамках исследования, во всех анализируемых примерах данный маркер выполняет именно эту функцию, так как с одной стороны он оформляется в отдельную синтагму или является словом с дополнительной просодической выделенностью, стоящим в начале высказывания, тем самым сигнализируя о переходе говорящего к обобщению ранее сказанной информации, как бы разграничивая то, что было сказано ранее и то, что будет произнесено после в *общем*. С другой стороны, в левом и правом контексте встречаются паузы хезитации или могут быть удлинения конечного звука как у слова в синтагме, стоящей перед в *общем*, так и в самом исследуемом маркере. Как кажется, это указывает на то, что говорящий в ходе преодоления возникшего коммуникативного затруднения подбирает нужные единицы для продолжения речи.

В примере (см. рис. 1) перед единицей в *общем* мы видим оборванную синтагму, удлинение но в ней, а также незаполненную паузу хезитации. После произнесения исследуемого маркера говорящий обобщает ранее сказанную им информацию.

Метод

С целью описания интонационных характеристик и жестикюляции прагматического маркера в *общем* на материале спонтанной устной речи был проведен акустический и слуховой анализ речи, а также анализ жестов, совершенных говорящими. В программе Wave Assistant была сделана транскрипционная расшифровка аудиоматериала и сегментация на фразы для исследования участков речевого сигнала, представляющих определенные супraseгментные особенности данного прагматического маркера. В программе ELAN для каждого видеофайла в первом слое была создана аннотация вербальной информации (про-

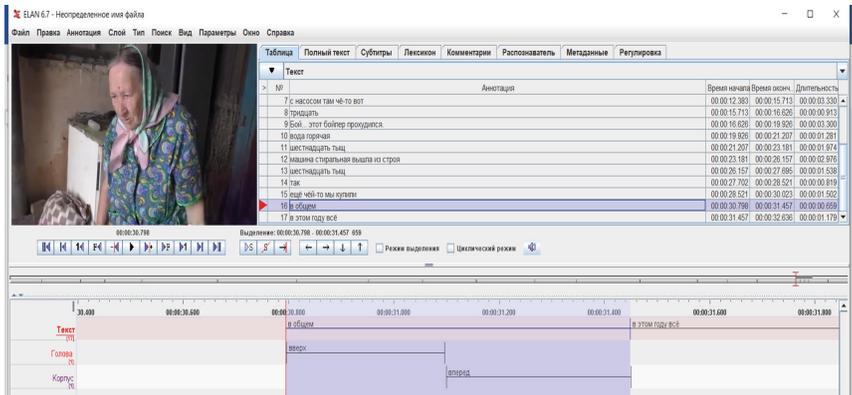


Рис. 2. Разметка речи и жестов в программе ELAN

изведено деление на синтагмы), а в последующих слоях сделана разметка невербальной информации таким образом, что каждому уровню соответствовал один жестикулятор (см. рис. 2). Затем для каждого контекста было описано сопоставление анализа интонационного оформления ПМ в *общем* и анализ жестикуляции, используемой при произнесении этой единицы.

Результаты и выводы

В ходе исследования интонационного оформления было выявлено, что в *общем* реализуется в трёх интонационных моделях, описанных Н. Б. Вольской:

- 17 употреблений (49%) реализуются по модели 06а (ИК-6 в классификации Е. А. Брызгуновой [2]), она характеризуется высоким ровным тоном (и на заударных гласных) с увеличением длительности гласных (см. рис. 3) [2];

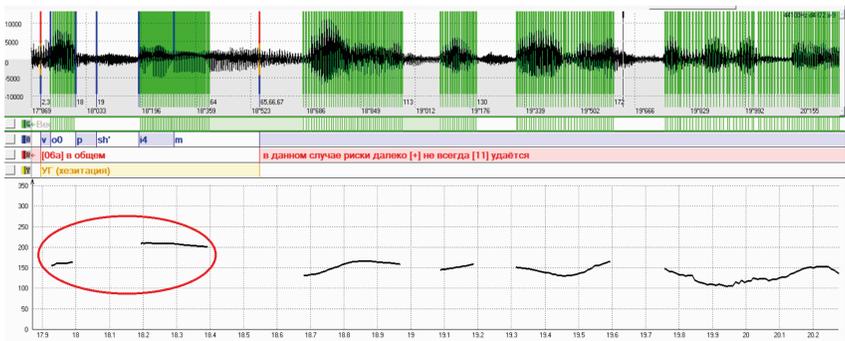


Рис. 3. Модель 06а, реализующаяся на единице в *общем*

- 15 примеров (43 %) произносятся с моделью 11, в которой происходит повышение тона на гласном интонационного центра с падением тона в постцентре по типу ИК-3 (см. рис. 4) [2, 3]

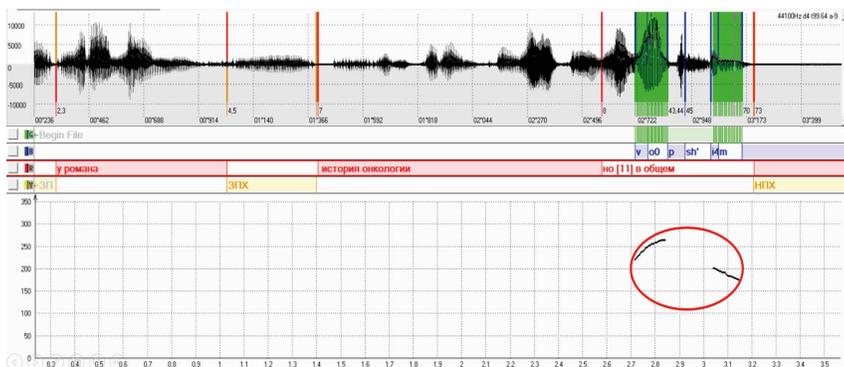


Рис. 4. Модель 11, реализующаяся на единице в *общем*

- 3 примера (8 %) реализуются по модели 02 (ИК-2), которая характеризуется нисходящей мелодикой с более высокого уровня с обязательным усилением удара на гласном (слоге) выделенного слова (см. рис. 5) [2, 3].

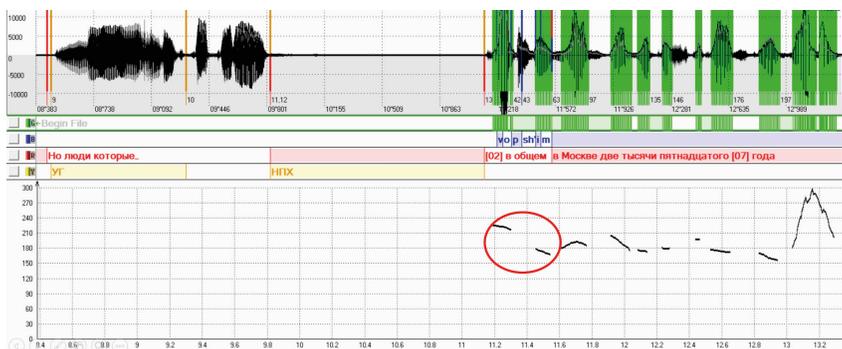


Рис. 5. Модель 02, реализующаяся на единице в *общем*

В процессе анализа визуального канала было определено, что в большинстве случаев (34 примера) данная единица сопровождается каким-либо жестом: движением головы, плечей, рук, глаз. В 24 случаях из 34 при произнесении ПМ в *общем* говорящие совершают движение вверх (поднимают голову, плечи, глаза, брови), дублируя тем самым движение тона на ударном слоге. В остальных случаях говорящие либо опускают голову (5 примеров), либо сначала опускают, а затем подни-

мают голову (5 примеров), что, вероятно, связано с тем, что они маркируют переход к новой синтагме, выделяя исследуемый ПМ как с помощью интонации, так и благодаря жестам.

Интересным наблюдением является то, что почти в половине случаев (16 случаев из 35) говорящие моргают в момент произнесения единицы *в общем*. Стоит предположить, что этот жест может также указывать на момент начала синтагмы.

Заключение

Исследование показало, что во всех примерах прагматический маркер *в общем* интонационно выделяется, эта единица сопровождается жестом, движение вверх каким-либо жестикулятором говорящих в 70 процентах случаев соотносится с восходящей интонацией, которую они реализуют при произнесении рассматриваемого маркера.

Данные результаты являются статистически значимыми, поскольку вероятность ошибки при отклонении от нулевой гипотезы (что интонация не соотносится с жестами) является меньше уровня значимости, который составляет 0,05 ($P\text{-value} = 0.0164$).

Полученные результаты позволяют предположить особое соотношение жестов и интонационного оформления на ПМ, будут использованы на дальнейших этапах анализа, и в будущем могут быть применены при разработке роботов-помощников (например, при обучении детей), а также в преподавании русского языка как иностранного.

Список использованной литературы

1. *Llopis Cardona A. et al. A Multimodal Approach to Catalan Pragmatic Markers: An Exploratory Study // Languages. 2024. № 9. С. 2.*
2. *Брызгунова Е. А. Интонационные конструкции // Русская фонетика. URL: <http://www.philol.msu.ru/~fonetica/index1.htm>*
3. *Вольская Н. Б., Скрелин П. А. Система интонационных моделей для автоматической интерпретации интонационного оформления высказывания: функциональные и перцептивные характеристики // Анализ разговорной русской речи: третий междисциплинарный семинар (АРЗ-2009) / сост. А. Л. Ронжин. СПб.: ГУАП, 2009. С. 28–40.*
4. *Гришина Е. А. Русская жестикуляция с лингвистической точки зрения. Корпусные исследования. М.: Изд. дом ЯСК, 2017. С. 27.*
5. *Прагматические маркеры русской повседневной речи: Словарь-монография / под ред. Н. В. Богданова-Бегларян. СПб.: Нестор-История, 2021. С. 61–66.*

АВТОМАТИЧЕСКОЕ ТРАНСКРИБИРОВАНИЕ РУССКОЙ УСТНОЙ РЕЧИ ПРИ ПОМОЩИ ИНСТРУМЕНТА KALDI¹

*Скоробагатько Лада Леонидовна
(Санкт-Петербургский государственный университет)*

Введение

Задача автоматического распознавания речи (Automatic Speech Recognition, или ASR) заключается в разработке алгоритмов и программ, которые способны преобразовывать аудиосигналы, содержащие естественную речь человека, в конвенциональный текстовый формат. Это может быть как задача бинарной классификации (например, необходимо указать, в каких аудиофрагментах содержится ответ «да», а в каких — «нет»), так и более сложные регрессионные задачи на непосредственную расшифровку текста [4]. Цель автоматического распознавания речи — облегчить и ускорить процесс обработки и анализа больших объемов аудиоданных, что может быть полезно в различных областях, таких как технологии голосового управления, транскрипция аудиозаписей, медицинская документация и др. Иначе говоря, результаты исследований в данной области широко применяются как в прикладных, так и в фундаментальных работах [1; 3; 7].

Кроме того, алгоритмы автоматического распознавания речи применяются при создании корпусов устной речи, основной задачей которых является подготовка материала для последующих экспериментальных и корпусных исследований процессов порождения и восприятия речи. Примером может быть Корпус русской устной речи [12]. Он используется в первую очередь для изучения восприятия русской устной речи. Корпус представляет собой собрание аудиофрагментов с их орфографической расшифровкой, а также транскрипцией, которая создается посредством слухового и инструментального анализа сигнала. Аннотирование аудиофрагментов выполняется вручную специально обученными экспертами с лингвистическим образованием. При этом происходит «последовательное прослушивание небольших (один слог) асемантических фрагментов записей спонтанных текстов с одновременным анализом соответствующих им динамических спектрограмм» (например, формант для аллофонов гласных) [13]. Таким образом, в ходе

¹ Научный руководитель — канд. филол. наук, доцент Е. И. Риехакайнен.

транскрибирования «фиксируется то, что было произнесено в речи на самом деле», то есть максимально объективно извлекаются акустические характеристики речи. Аудиофрагмент при этом предварительно сегментируется на т. н. межпаузальные интервалы, т. е. отрывки речи между двумя паузами (паузой при этом признается отсутствие полезного речевого сигнала продолжительностью не менее 100 мс) [10].

Ранее уже предпринимались попытки автоматизировать ряд процессов, связанных с анализом материалов корпуса. Для выявления признаков, способствующих редукции словоформ, применялись методы машинного обучения, например, случайный лес, сверхслучайные деревья и логистическая регрессия. Было выявлено, что на выпадение звуков в словоформах влияют такие факторы, как количество слогов, а также отнесенность к определенным частям речи (в предложениях, прилагательных и наречиях редукция происходит значительно чаще) [11].

Тем не менее, пополнение корпуса — в силу необходимости осуществления акустико-фонетической транскрипции — до сих пор требует достаточно больших вложений человеческого и временного ресурса. Ускорить эти процессы может использование алгоритмов автоматического распознавания речи, задачей которого будет являться не орфографическая расшифровка, а транскрибирование устной речи по особой конвенции. Для этого необходим поиск оптимального метода или инструмента для создания модели.

Для обучения моделей решения задачи ASR существует инструмент Kaldi [5], исходный код которого написан на языке C++. Некоторые уже обученные на различных данных модели находятся в открытом доступе, и можно как использовать их, так и обучить новый алгоритм на собственных данных.

Создание моделей в Kaldi строится на лингвистическом и акустическом моделировании. Первое учитывает информацию с различных языковых уровней и контекст, что помогает повысить точность распознавания звукового фрагмента; второе фокусируется на извлечении акустических параметров речевого сигнала [8].

Алгоритмы Kaldi были опробованы на материале нескольких языков. Например, он показал хорошие результаты для швейцарского диалекта немецкого языка [6], а также для мандаринского диалекта китайского языка [2], однако внимание в этих работах в основном уделялось лингвистическому моделированию.

Таким образом, в нашем исследовании для автоматического транскрибирования устной речи, ввиду перечисленных выше достоинств, решено было использовать именно инструмент Kaldi.

Материал

В качестве материала исследования были использованы данные Корпуса русской устной речи [12], а именно 3 аудиофайла в формате wav общей продолжительностью 81 минута и 40 секунд, поделенные на межаузуальные интервалы. Количество отрезков составило 3083 для всех файлов. Из данных были исключены хезитации, неразборчивая речь, промежутки с одновременной речью нескольких дикторов, гортанные смывки, участки ларингализации, а также паузы, поскольку нас в первую очередь интересовало распознавание полезного речевого сигнала.

Данные были случайно разбиты на обучающую и тестовую выборки в соотношении 8 к 2; таким образом, в обучающую вошли 2466 межаузуальных интервалов, а в тестовую — 617.

Метод

В Kaldi существует возможность обучать модели несколькими способами, включая монофонный и трифонный.

Монофонный способ — это обучение акустических моделей на основе отдельных фонем, и алгоритм учится распознавать акустические характеристики каждого звука изолированно друг от друга. Этот подход относительно прост и требует меньше вычислительных ресурсов.

Трифонный способ, напротив, включает в себя обучение акустических моделей на основе последовательностей из трех фонем. Учитывается контекст, в котором находится звук, то есть положение аллофона. Трифонное обучение, как правило, показывает более высокие результаты, чем монофонное, однако оно более затратное с точки зрения вычислительных ресурсов.

Для оценки качества моделей использовалась метрика Word Error Rate (далее — WER). Это отношение ошибок при распознавании к общему количеству слов; под ошибками в данном случае подразумеваются пропуски слов, вставки новых, а также замены внутри самого слова. Чем ниже WER, тем выше качество модели [9].

Самую первую модель было решено обучить с большим словарем, в котором были собраны все слова, которые могли потенциально встретиться в обучающей и тестовой выборках. В результате показатель метрики WER составил 88,85 и 89,92 для монофонного и трифонного обучения соответственно.

Из-за низкого качества результатов, а также из-за того, что в таком случае алгоритм пытается распознать слово целиком, а не отдельные звуки, было принято решение изменить способ обучения, включив

в словарь исключительно отдельные аллофоны русского языка. При этом мы опирались на систему транскрипции, которая используется в Корпусе русской устной речи и ориентирована на выделение тех звуков, которые слушающий различает в процессе восприятия речи [14]. Объем словаря таким образом составил 95 токенов.

Кроме того, были проведены эксперименты с увеличением длины N-грамм, поскольку было сделано предположение, что при большем количестве учитываемого контекста качество работы модели будет улучшаться. Таким образом, всего на словаре из 95 токенов было обучено 15 разных моделей, отличавшихся только длиной N-грамм.

Результаты и выводы

Базово при уменьшении размерности словаря показатели метрики WER (под которой в данном случае — в силу особенностей организации словаря — понимается количество ошибок при распознавании отдельных аллофонов) улучшились примерно в 2 раза в сравнении с предыдущей моделью, где присутствовал словарь большей размерности: 44,68 и 39,86 для монофонного и трифонного обучения соответственно. При увеличении длины N-грамм качество результатов также росло, особенно значимо при $N = 6$ и $N = 4$. Это может быть связано с тем, что усредненная длина высказываний лежит в промежутке этих значений (средняя длина межпаузальных интервалов составляет около 17 аллофонов). В целом можно предположить, что увеличение длины N-грамм влияет на количество ошибок (при увеличении длины оно снижается), однако при слишком больших значениях N улучшения не происходит.

Таким образом, в ходе работы над моделью были сделаны следующие выводы:

- 1) модели с большим словарем, токенами которых являются слова как промежутки между двумя пробелами в орфографической расшифровке, показывают достаточно большой процент ошибок, а также не отвечают исходной задаче создания автоматической акустико-фонетической транскрипции;
- 2) модели с небольшими словарями, токенами которых выступают отдельные аллофоны, напротив, лучше справляются с поставленной задачей;
- 3) увеличение длины N-грамм до значений от 4 до 6 значительно снижает WER; дальнейшее увеличение длины N-грамм не приводит к значительному улучшению результатов;
- 4) в среднем трифонное обучение показывает более высокие результаты, чем монофонное.

Заключение

Kaldi показал себя как эффективный инструмент для обучения моделей автоматического распознавания русской устной речи. В дальнейшей работе планируется улучшить существующие алгоритмы, а именно:

- 1) протестировать другие способы обучения помимо трифонов и монофонов;
- 2) проверить, повлияет ли фактор указания спикера на качество распознавания аллофонов;
- 3) возможно, сбалансировать обучающую и тестовую выборки по другим критериям, чтобы повысить эффективность обучения;
- 4) проанализировать более подробно, в каких случаях возникают ошибки (например, какие аллофоны чаще всего путаются друг с другом, какие опускаются, какие распознаются лучше всего и т. п.).

Благодарности

Работа выполнена при поддержке СПбГУ, шифр проекта 103923108.

Список использованной литературы

1. *Brown C., Russel M. J., Series R., Skilling A.* Applications of automatic speech recognition to speech and language development in young children // Proceedings of the 4th International Conference on Spoken Language. Philadelphia, 1996. P. 1–5.
2. *Furui S.* Automatic Speech Recognition and Its Application to Information Extraction // Department of Computer Science Tokyo institute of Technology. Tokyo, 2015. P. 11–20.
3. *Bu H., Du J., Na X., Wu B., Zheng H.* AISHELL-1: An Open-Source Mandarin Speech Corpus and A Speech Recognition Baseline. <https://arxiv.org/abs/1709.05522>.
4. *Furui S.* Automatic Speech Recognition and Its Application to Information Extraction // Department of Computer Science Tokyo institute of Technology. Tokyo, 2015. P. 11–20.
5. *Jurafsky D., Martin J. H.* Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. URL: <https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf>
6. Kaldi: официальный сайт. URL: <https://kaldi-asr.org/>
7. *Kew T., Nigmatulina I., Samardzic T.* UZH TILT: A Kaldi recipe for Swiss German Speech to Standard German Text // Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects. Barcelona, 2020. P. 15–24.

8. *Nigmatulina I., Zuluaga-Gomez J., Prasad A., Sarfjoo M.M., Motlicek P.* A two-step approach to leverage contextual data: speech recognition in air-traffic communications // *Proceedings of International Conference on Acoustics, Speech and Signal Processing*. Singapore, 2022. P. 6282–6286.
9. *Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hanneman M., Motlíček P., Qian Y., Schwarz P., Silovsky J., Stemmer G., Vesel K.* The Kaldi speech recognition toolkit. URL: https://www.danielpovey.com/files/2011_asru_kaldi.pdf
10. *Tomás J., Àngel Mas J., Casacuberta F.* A Quantitative Method for Machine Translation Evaluation // *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in “Natural Language Processing: are evaluation methods, metrics and resources reusable?”*. Columbus, 2008. P. 27–34.
11. *Венцов А.В., Нигматулина Ю.О., Раева О.В., Риехакайнен Е.И., Слепкокурова Н.А.* Корпус русских спонтанных текстов: структура и единицы // *Труды международной конференции «Корпусная лингвистика–2013»*. СПб., 2013. С. 223–231.
12. *Дайтер М.В., Риехакайнен Е.И.* Использование методов машинного обучения для изучения фонетической редукции слов в русской устной речи // *Международная конференция «Лингвистический форум 2020: Язык и искусственный интеллект»*. М., 2020. С. 79–80.
13. Корпус русской устной речи. URL: <https://russpeech.spbu.ru/>
14. *Нигматулина Ю.О., Раева О.В.* Исследование русской спонтанной речи: новые методы — новые результаты // *Язык и метод: Русский язык в лингвистических исследованиях XXI века*. 2016. С. 373–382.
15. Транскрипция Корпуса русской устной речи. URL: <https://russpeech.spbu.ru/transkrip.htm>

ОШИБКИ В ПРОЧТЕНИИ СЛОВ-ОМОГРАФОВ И НЕКОТОРЫЕ ДРУГИЕ ФАКТОРЫ В ВОСПРИЯТИИ СИНТЕЗИРОВАННОЙ РЕЧИ¹

Столяров Иван Игоревич

(Санкт-Петербургский государственный университет)

Введение

Вследствие стремительного развития технологий синтеза речи становятся как никогда актуальными вопросы восприятия синтезированной речи человеком и оценки качества систем синтеза. Традиционно оцениваются такие глобальные аспекты синтезированной речи, как ее разборчивость и естественность [8; 9; 11]. Для оценки естественности речи можно рассматривать качество различных этапов синтеза — лингвистической обработки исходного текста, дальнейшей просодической обработки и т. д. [7]. Важным компонентом лингвистической обработки текста является снятие омографии, т. е. выбор одного из нескольких вариантов произнесения слова-омографа. В русском языке омография может быть связана с разной постановкой ударения («замок» — *за́мок* или *замо́к*) и/или неразличением буквы «ё» на письме («все» — *все* или *всё*; «берег» — *бе́рег* или *берёг*).

Представляется, что в настоящее время технологии синтеза речи на основе нейронных сетей достигли больших успехов в порождении не только разборчивой, но и естественно звучащей речи (с точки зрения интонации, ритма и т. д.) [1; 2]. В этой связи возникает вопрос — продолжает ли правильное прочтение отдельных слов-омографов играть существенную роль в оценке качества синтезированной речи?

Цель настоящей работы — оценить, насколько ошибки в произнесении омографов влияют на качество синтезированной речи, сгенерированной с помощью систем синтеза последнего поколения. В рамках исследования был проведен перцептивный эксперимент и проанализированы его результаты. Помимо выводов о влиянии неправильного прочтения омографов на качество синтеза, представлены также наблюдения о роли других факторов в восприятии синтезированной речи.

¹ Научный руководитель — доцент СПбГУ, канд. филол. наук О. А. Митрофанова.

Материал

В эксперименте использовалось 24 аудиозаписи длительностью от 8 до 15 секунд. Записи были разделены на три равные по количеству группы:

- 8 записей естественной речи (для удобства и краткости обозначим эту группу «группа Е»);
- 8 записей синтезированной речи, не содержащие ошибок в прочтении омографов («группа С+»);
- 8 записей синтезированной речи, содержащие ошибки в прочтении омографов («группа С–»). В 4 записях ошибки были обусловлены исключительно неправильной постановкой ударения, в 4 других — связаны с неправильным прочтением буквы «ё».

Каждая запись являлась монологическим высказыванием, состоящим из 1–3 предложений. Образцы естественной речи представляли собой записи 8 различных дикторов из корпуса профессионального чтения CORPRES, разработанного на Кафедре фонетики и методики преподавания иностранных языков СПбГУ [10]. Для записей групп С+ и С– озвученные примеры были взяты из Национального корпуса русского языка [3]. В таблице 1 приведены примеры текстов, использованных в записях разных групп.

Таблица 1. Примеры текстов записей из разных групп

Группа записей	Примеры озвученных текстов	Комментарии
Группа Е	<i>Чаще всего Интернетом пользуются молодые и образованные россияне, жители крупных городов, а значит, что всемирная паутина имеет большие перспективы развития в нашей стране.</i>	запись gte0068 из корпуса CORPRES
Группа С+	<i>Дети выражают свои эмоции не так, как взрослые. Если взрослый может выплакать свое горе и таким образом дать ему выход, то детям свойственна скрытая форма переживания.</i>	
Группа С–	<i>Один раз в Португалии у меня был сильный кашель, но пришлось работать. Это была мука́. Вообще, героизм ни к чему, только негативно на качестве работы сказывается.</i>	<i>мука́</i> вместо <i>му́ка</i>
	<i>Приходят в упадок традиционные виды промыслов, ремесел, культура и искусство. В романе описано одно из таких сел с красивым названием Снежное.</i>	<i>сел</i> вместо <i>се́л</i>

Группа записей	Примеры озвученных текстов	Комментарии
Группа С–	<i>Первым делом мы его успокоили. Это нужно было прекратить немедленно, ведь только каким-то чудом сестры до сих пор не обратили внимания на его побитую рожу.</i>	<i>сестры</i> вместо <i>сёстры</i>

Для синтеза было решено использовать демоверсии систем SaluteSpeech и Yandex SpeechKit, находящиеся в открытом доступе (для каждой записи использовался уникальный голос) [4; 5]. Во всех группах соблюдался паритет между мужскими и женскими голосами. При проведении эксперимента записи трех групп были перемешаны.

Методика

Для создания анкеты с аудиозаписями и сопровождающими их вопросами использовалась программа Google Forms. Каждой записи соответствовало три «страницы» анкеты. На первой странице респондент должен был прослушать запись дважды (через паузу в 4 секунды) и ответить на вопрос: «Как вы считаете, это естественная или синтезированная речь?». Если респондент выбирал ответ «Естественная речь», ему предлагалось сразу перейти на третью страницу; если был выбран ответ «Синтезированная речь», то переход осуществлялся на вторую страницу, где испытуемый должен был оценить качество синтезированной речи от 1 до 10 по ряду критериев:

- расстановка и длительность пауз;
- расстановка ударений;
- интонационное оформление;
- темп и ритм;
- общее качество речи, отсутствие ошибок в произнесении отдельных звуков.

После выставления оценок респондент переходил на третью страницу. На третьей (последней) странице испытуемому предлагалось ответить в произвольной форме на вопрос «Благодаря чему вы определили, что это естественная/синтезированная речь?» или оставить этот вопрос без ответа.

Выбор критериев для оценки качества синтезированной речи опирался на методику, предложенную в работе [6]. Ожидалось, что ошибки в прочтении омографов повлияют на оценки по критериям «Расстанов-

ка ударений» (*мука́* вместо *му́ка*, *сестры́* вместо *сёстры*) и «Общее качество речи, отсутствие ошибок в произнесении отдельных звуков» (*все* вместо *всё*, *сел* вместо *сёл*).

Результаты

В эксперименте приняли участие 42 респондента. Все респонденты имели высшее образование, некоторые из них являлись профессиональными лингвистами и фонетистами. Половой и возрастной состав участников эксперимента не исследовался.

Результаты эксперимента показали, что ошибки в прочтении омографов действительно имеют большое влияние на восприятие речи как синтезированной: если для записей группы С+ респонденты в 18,9% случаев выбирали ответ «Естественная речь», то для записей группы С– этот показатель примерно в три раза ниже — 6,7% (см. табл. 2). Отметим также, что в некоторых случаях респонденты воспринимали даже записи группы Е как синтезированную речь (13,9%): такой результат, вероятно, объясняется психологическим эффектом (испытуемые были настроены искать признаки «синтезированности» даже в тех записях, где они заведомо отсутствовали) и косвенно отражает сложившиеся представления о высоком качестве синтеза на современном этапе.

Таблица 2. Результаты оценивания записей разных групп как примеров естественной/синтезированной речи

Группа записей	Процент респондентов, оценивших записи как естественную речь	Процент респондентов, оценивших записи как синтезированную речь
Группа Е	86,1	13,9
Группа С+	18,9	81,1
Группа С–	6,7	93,3

Интересно отметить, что по сравнению с группой С+ записи группы С– получили в среднем более низкие оценки не только по критериям «Расстановка ударений» и «Общее качество речи, отсутствие ошибок в произнесении отдельных звуков» (что было ожидаемо), но и по всем остальным критериям, никак не связанным со снятием омографии (см. рис. 1). В этой связи можно предположить, что ошибки в прочтении омографов не только «выдают» синтезированную речь, но и ухудшают общее впечатление от ее прослушивания.

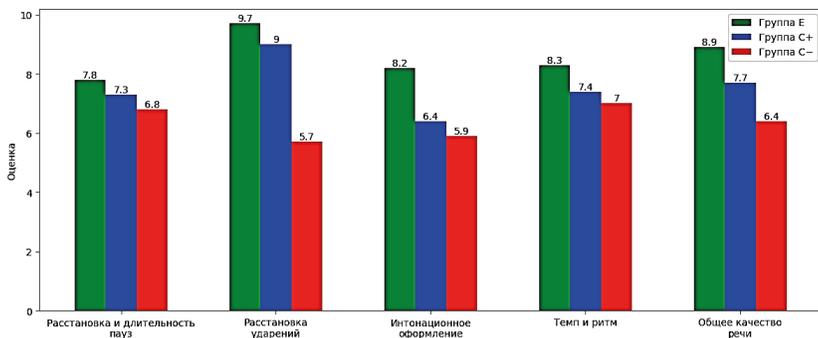


Рис. 1. Усредненные оценки записей разных групп по пяти критериям

Также были вручную обработаны комментарии респондентов, ответивших на вопрос «Благодаря чему вы определили, что это естественная/синтезированная речь?». Так, на основе комментариев к записям C+ и C- были выделены группы факторов (ошибки, недочеты и т. д.), из-за которых, по мнению респондентов, речь воспринималась как синтезированная. Для каждой группы факторов было подсчитано число соответствующих упоминаний в комментариях (см. табл. 3).

В группе C- респонденты чаще всего упоминали ошибки в прочтении омографов (166 комментариев). Отметим, что в нескольких записях этой группы, которые в целом получили высокие средние оценки по разным критериям, именно неправильно произнесенные омографы убедили некоторых респондентов, что звучит все-таки не естественная речь — см., например, комментарии: «запись шикарная, я меняла мнение лишь на неверном произношении на слове “сел” (должно быть “сёл”)», «как будто специально сказали “мука” неправильно, а звучит как человек» и т. д. Также среди недочетов синтезированной речи участники экспериментов выделили интонацию (122 и 77 комментариев для групп C+ и C- соответственно), темп и ритм (30 и 28 комментариев), паузацию (25 и 29 комментариев).

Для записей группы E были, наоборот, проанализированы комментарии, в которых респонденты описывали, что именно помогло им сделать выбор в пользу естественной речи. Удалось выявить следующий группы факторов, которые представлены в таблице 4.

Таким образом, можно увидеть, что для большинства респондентов живая речь человека характеризуется прежде всего правильными интонациями (54 комментария), естественностью некоторых экстралингвистических характеристик речи (в первую очередь дыхания — 50 комментариев) и адекватной паузацией (35 комментариев).

Таблица 3. Результаты анализа комментариев к записям групп С+ и С–

Факторы, влияющие на восприятие речи как синтезированной	Примеры комментариев	Кол-во коммент. к записям гр. С+	Кол-во коммент. к записям гр. С–
Ошибки в прочтении омографов	<i>слово «всё» произнесено как «все» крайне неестественная ошибка в ударении</i>	0	166
Неправильная интонация	<i>интонация всегда вверх, немного странно звучит не очень естественная интонация в конце фразы</i>	122	77
Неестественный темп и ритм	<i>очень быстро произносятся слова нарушение ритмичности речи</i>	30	28
Неестественные паузы	<i>слишком большие паузы странные паузы в середине фразы</i>	25	29
Шумы и искажения	<i>шум в голосе, странные призвуки в голосе слышны «потрескивания»</i>	21	23
Неестественный тембр голоса	<i>что-то металлическое в голосе роботизированный голос</i>	13	19
Сегментные ошибки	<i>некоторые гласные оборваны окончания слишком быстро произносятся</i>	8	9
Неестественные экстралингвистические особенности речи (дыхание и т. д.)	<i>очень странные вдохи после каждого второго слова дыхание не к месту</i>	4	5
Неуместная эмоциональность/ безэмоциональность речи	<i>речь звучит слишком «восторженно» безэмоциональность, хотя по контексту она может предполагаться</i>	2	3

Стоит отметить, что если среди недочетов синтезированной речи респонденты упоминали механистичность и однообразие некоторых параметров речи («одинаковые паузы», «стандартизованность

Таблица 4. Результаты анализа комментариев к записям группы Е

Факторы, влияющие на восприятие речи как естественной	Примеры комментариев	Кол-во коммент. к записям гр. Е
Правильная интонация	<i>благодаря правильному интонированию интонационно подчеркнуты некоторые смысловые компоненты</i>	54
Естественные экстралингвистические особенности речи (дыхание и т.д.)	<i>здесь показалось, что речь более соответствует естественному дыханию человека слышен звук слюней во рту в какой-то момент</i>	50
Естественные паузы	<i>более длинные естественные паузы различная длительность пауз после произнесения слов, что придает некоторую эмоциональную окраску сказанному</i>	35
Естественный темп	<i>естественный темп речи живой темп</i>	17
Естественный тембр голоса	<i>в голосе чувствуется глубина</i>	17
Индивидуальные особенности диктора	<i>чуть картавое r интересный прононс</i>	7

пауз», «очень «техничная», слишком ровная интонация», «смутили интонации — вся речь произнесена абсолютно ровно, с одной интонацией» и т.д.), то для естественной речи, наоборот, отмечались вариативность и разнообразие тех же параметров («более свободные паузы», «вольные паузы», «более широкий диапазон интонации», «ярко выраженная интонация» и т.д.).

Большой интерес представляет также восприятие экстралингвистических особенностей речи. Очевидно, что, например, звуки вдоха и выдоха придают речи естественность — см. многочисленные комментарии к записям группы Е вроде «слышно дыхание человека», «вдох перед произнесением слов», «слышен вдох на конце (если это уловка, то работает супер)» и т.д. В системе синтеза Yandex SpeechKit к речи нескольких дикторов автоматически добавляются звуки дыхания, однако они не всегда звучат уместно — так, респонденты отмечают: «речь не соответствует естественному дыханию», «очень странные вдохи по-

сле каждого второго слова, вряд ли человеку нужно вдыхать настолько часто» (ср. с комментариями к записям группы Е — «расстановка пауз и темп под естественное дыхание человека», «речь более соответствует естественному дыханию человека»).

Заключение

Таким образом, результаты исследования показывают, что правильность прочтения омографов сохраняет важность для оценки качества синтеза речи даже при достижении высоких результатов по другим критериям оценивания: во-первых, ошибки в прочтении омографов дают слушателям ясно понять, что звучит синтезированная, а не естественная речь; во-вторых, при наличии таких ошибок слушатели имеют тенденцию ниже оценивать качество синтеза по различным критериям, даже по тем, которые не связаны с расстановкой ударений и/или прочтением буквы «ё». Также в исследовании делается вывод о том, что на восприятие речи как естественной или синтезированной больше всего влияют такие факторы, как интонация, экстралингвистические характеристики речи (в первую очередь дыхание), паузация, темп и ритм.

Список использованной литературы

1. Калиев А., Рыбин С. В. Синтез речи: прошлое и настоящее // Компьютерные инструменты в образовании. 2019. № 1. С. 5–28.
2. Киреев Н. С., Ильющин Е. А. Обзор существующих алгоритмов преобразования текста в речь // International Journal of Open Information Technologies. 2020. № 7. С. 84–90.
3. Национальный корпус русского языка. URL: <https://ruscorpora.ru/> (дата обращения: 10.02.2024).
4. Попробуйте технологию SaluteSpeech прямо сейчас. URL: <https://developers.sber.ru/portal/products/smartspeech> (дата обращения: 18.02.2024).
5. Попробуйте Yandex SpeechKit API. URL: <https://yandex.cloud/ru/services/speechkit#demo> (дата обращения: 18.02.2024).
6. Соломенник А. И. Оценка качества селективного синтеза речи: методы и результаты. Дис. ... канд. филол. наук. МГУ им. М. В. Ломоносова. Филологический факультет. М., 2016.
7. Соломенник А. И., Таланов А. О., Соломенник М. В., Хомицевич О. Г., Чистиков П. Г. Оценка качества синтезированной речи: проблемы и решения // Изв. вузов. Приборостроение. Тематический выпуск «Речевые информационные системы». 2013(а). № 2. С. 38–42.
8. Фролова И. Г. Естественность и разборчивость как параметры оценки качества синтезированной речи // Экспериментальная фонетика. М.: Изд-во МГУ, 1989. С. 95–102.

9. *Loizou P. C.* Speech Quality Assessment // *Multimedia Analysis, Processing and Communications* / eds Lin Weisi et al. Springer, 2011. P.623–654.
10. *Skrelin P., Volskaya N., Kocharov D., Evgrafova K., Glotova O., Evdokimova V.* A Fully Annotated Corpus of Russian Speech // *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA). 2010. P.109–112.
11. *Taylor P.* Text-to-Speech Synthesis. Cambridge: Cambridge University Press, 2009. 474 p.

ПРОСОДИЧЕСКИЕ ХАРАКТЕРИСТИКИ НАЧАЛА ЛЕКЦИИ И ИХ ВОСПРИЯТИЕ¹

Титова Рената Руслановна
(Санкт-Петербургский государственный университет)

Введение

В современных научных работах говорится о том, что голос преподавателя может влиять на усвоение материала, поскольку интонация преподавателя (соответствующая высота тона, громкость и их вариации) может помочь учащемуся определить основные моменты речи. Предпочтительный голос для преподавания считается приятным, мотивирующим и способным привлечь внимание, и воспринимается как нейтральный, с низким тембром и медленным темпом [1].

В пособиях по методике преподавания, как правило, недостаточно подробно с фонетической точки зрения описывается использование голоса как инструмента достижения цели в обучении. Используются такие описания, как: «добрый голос», «звонкий голос», «злобно-строгий голос», «низкий голос» [5] и, говоря о голосе, чаще всего подразумевают мелодические характеристики.

Интонацию определяют как комплекс характеристик, которые формируют целостную систему звуковых средств, взаимодействуя с синтаксическим строением и лексическим составом высказывания [3].

Целью исследования является определение значимости определенных просодических средств в разных структурах лекционных и практических занятий в высших учебных заведениях. Выявление влияния интонационных характеристик на их восприятие обучающимися может помочь при составлении методических рекомендаций для преподавателей, голосовых помощников, при использовании системы искусственного интеллекта в образовательных целях.

Поскольку перцептивная релевантность отдельных просодических параметров речи преподавателя до сих пор изучена недостаточно, в данном исследовании проводится перцептивный анализ просодических характеристик лекции до и после внесения модификаций в звуковой сигнал.

¹ Научный руководитель — доцент СПбГУ, канд. филол. наук У.Е. Кочеткова.

Материал

Материалом данного исследования послужили две записи лекций в высших учебных заведениях (мужской и женский голоса) в городе Санкт-Петербург и городе Барнаул. Первый лектор (далее «1 лектор») — мужчина, с трехлетним стажем преподавания, второй лектор (далее «2 лектор») — женщина с десятилетним стажем преподавания. Обучающаяся аудитория: первый и четвертый курс соответственно. Направление обучения у обеих аудиторий являлось гуманитарным.

Аудиозаписи лекций были разделены на структурные элементы (начало, середина, конец)

1. Для первых компонентов (начало) в программе WaveAssistant был проведен анализ просодических характеристик: темпа речи, диапазона основного тона, интонационных конструкций (ИК определялись по системе Е. А. Брызгуновой [2], интонационные модели — по системе Н. Б. Вольской [4]), типов пауз.
2. Проведен пилотный перцептивный эксперимент, участниками которого стали 46 женщин и 10 мужчин. Они оценивали нравятся ли им звучание голоса лектора, заинтересовал ли он их, какие эмоции они слышат в его речи.
3. Затем эти результаты соотносились с акустическим анализом записей.

На первом этапе в перцептивном эксперименте было представлено 2 отрывка обращения «Друзья, ...» как отдельной синтагмы и обращение (вырезанного слова «друзья» из высказывания «Так друзья, давайте, я думаю, мы с вами потихоньку начнем...»). Респонденты должны были выбрать чья реализация им понравилась больше. Они оценивали отрывок как «приятный». Затем был представлен полный отрывок одного лектора (затем второго), и респонденты оценивали был ли приятен им этот голос, заинтересованность, эмоциональную выразительность, стиль преподавания и хотели ли бы они пойти на обучение к преподавателю с таким голосом. В качестве классификации эмоций для определения эмоциональной выразительности была взята классификация П. В. Симонова (1966) «Деление эмоций по виду контакта живых существ». Далее производилась модификация вступительной части лекции (удаление всех пауз из сигнала) и участники оценивали изменилось ли их восприятие.

Результаты и выводы

Проведенный сравнительный анализ позволил выявить общие различия в речи лекторов. Особую трудность представляет отнесение пауз к определенному типу, так как пауза может сочетать в себе несколько функций одновременно. Например, выделение и хезитацию или вдох и хезитацию. Речь 2 лектора характеризуется меньшей длительностью по времени, разнообразием и длительностью пауз, меньшим количеством слов в синтагме, членением фразы на короткие интонационные группы, как правило, равные одной словоформе. Средняя частота основного тона и его диапазон ожидаемо больше у 2 лектора, так как это женский голос (рис. 1).

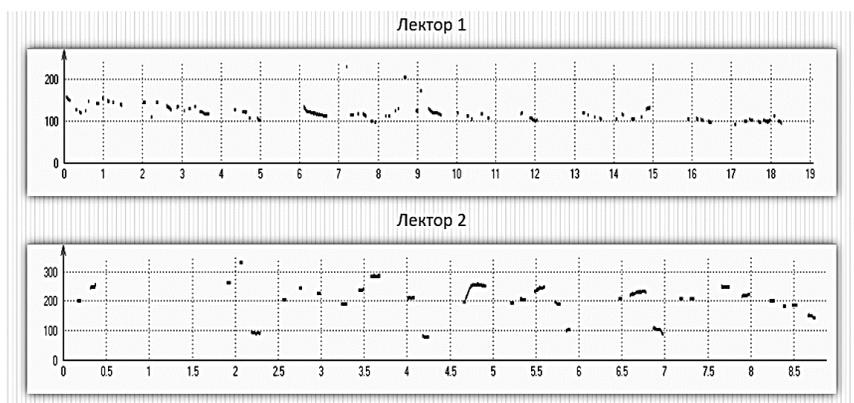


Рис. 1. Графики основного тона лекторов

У 1 лектора было выявлено нетипичное мелодическое оформление при синтагматическом членении, наличие такого явления, как «удлинение гласных» при паузах вдоха и незаполненных паузах хезитации, а также присутствие ларингализации. Средний артикуляторный темп речи так же выше у 2 лектора. Светлана Борисовна Степанова в своей статье приводит данные, полученные на материале Звукового корпуса русского языка (ЗКРЯ), о среднем артикуляторном темпе носителей русского языка, который оказался равным 5,31 сл/с [6]. Это опровергает некоторые выводы, которые приводит Виктория Андреевна Метлова в своей статье: «Темп речи в свободной коммуникации: социолингвистический аспект» о том, что «иногда в официальной ситуации темп речи меняется, становится медленнее, в неофициальной (дома) или же на работе, но с друзьями он становится быстрее». В данном исследовании мы можем наблюдать, что темп речи обоих преподавателей больше или равен

среднему темпу русскоговорящего человека, при том, что они оба находились в рабочих условиях. Результаты представлены в таблицах 1 и 2.

Таблица 1. Сравнительная характеристика

	Лектор 1	Лектор 2
Пол:	М	ж
Общая длит-ть начала:	19,098 сек	8,885 сек
Длительность пауз:	5,8 сек	2,9 сек
Средняя длит-ть пауз:	0,531 сек	0,347 сек
Доля пауз:	30 %	33 %
Удлиненные гласные:	4	0
Ларингализция:	+	-
Среднее кол-во слов в синтагме:	3,5	1,5
Средн. ОТ:	120.492 Гц	215.356 Гц
Диапазон ОТ:	123 Гц	248 Гц
Макс. Размах подъема/спада:	4 Гц / 17 Гц	70 Гц / 14 Гц
Средний темп речи:	5,3 сл/с	5,6 сл/с

Таблица 2. Сравнительная характеристика пауз

	Лектор 1	Лектор 2
Пол:	м	ж
Общая длит-ть начала:	19,098 сек	8,885 сек
Длительность пауз:	5,8 сек	2,9 сек
Средняя длит-ть пауз:	0,531 сек	0,347 сек
Доля пауз:	30 %	33 %
Виды пауз:	НПХ, ПВД, ЗПХ, ПС, ПМ	ПМ (ПС, ПВ)
Смысловые (ПС)	1	2
Паузы выделения (ПВ)	0	6
Паузы молчания (ПМ)	3	8
Средняя длит-ть:	0,169 сек	0,347 сек
Паузы хезитации (НПХ/ ЗПХ)	2/2	0
Паузы вдоха/выдоха (ПВД)	2	0
Пауз на границах синтагм:	6 из 8	8 из 8

Также благодаря анализу просодических характеристик было обнаружено преобладание неконечных синтагм, использование ИК-3, по типологии Е. А. Брызгуновой [2] в речи обоих лекторов. Первую и последнюю синтагмы оба лектора оформляли ИК-2.

Результаты перцептивного эксперимента выявили отсутствие разницы между восприятием обращения к аудитории «Друзья, ...» как отдельной синтагмы и обращением внутри синтагмы (вырезанного слова «друзья» из высказывания «Так друзья, давайте, я думаю, мы с вами потихоньку начнем...»). При модификации вступительной части лекции (удаление всех пауз из сигнала) большинство участников отметили отсутствие изменений восприятия, однако 28 % респондентов отметили улучшение восприятия при прослушивании в речи 1 лектора и 26 % — в речи 2 лектора. У 2 лектора участники опроса выделили более положительные эмоции: уверенность, спокойствие, радость, удовольствие, чем у 1 лектора: спокойствие, беспокойство, безразличие, уверенность. Оба голоса большинством были оценены как «приятные», несмотря на различие акустических характеристик. t-критерий Стьюдента по всем выборкам показал, что средние значения оценок голоса для лектора 1 и лектора 2 одинаковы.

При этом, суммируя все результаты опроса, более высокие показатели (голос оценивался как более приятный/было интересно/хотели бы обучаться у данного лектора) получил лектор 2. Эти наблюдения подчеркивают важность интонационного оформления в лекциях, которые могут влиять на эффективность обучения и восприятие аудиторий.

Список использованной литературы

1. Talking Like a Teacher — A Study of Pre-Service Teachers' Voice and Speech Characteristics in Learning and Teaching Situations Kati Järvinen et al. URL: https://www.researchgate.net/publication/378330577_Talking_Like_a_Teacher-A_Study_of_Pre-Service_Teachers'_Voice_and_Speech_Characteristics_in_Learning_and_Teaching_Situations
2. Брызгунова Е. А. Звуки и интонации русской речи. М., 1981.
3. Брызгунова Е. А. Русская грамматика. М.: Наука, 1980.
4. Вольская Н. Б., Скрелин П. А. Система интонационных моделей для автоматической интерпретации интонационного оформления высказывания: функциональные и перцептивные характеристики // Третий междисциплинарный семинар «Анализ разговорной русской речи» (Санкт-Петербург, 104 26–27 августа 2009 г.). СПб.: С.-Петерб. гос. ун-т, 2009. С. 28–40.
5. Ершов П. М., Ершова А. П., Букатов В. М. Общение на уроке, или Режиссура поведения учителя. 2-е изд., перераб. и доп. М.: Московский психолого-социальный институт; Флинта, 1998. 366 стр.

6. *Степанова С. Б.* Темп спонтанной русской речи (по материалам Звукового корпуса повседневного общения «Один речевой день») // *Язык и речевая деятельность*. СПб., 2010–2011. Т. II. С. 204–214.

АНИМАЦИОННЫЕ ФИЛЬМЫ НА ЗАНЯТИЯХ ПО АНГЛИЙСКОЙ ФОНЕТИКЕ¹

*Филимонова Надежда Александровна
(Санкт-Петербургский государственный университет)*

Введение

Цель данного исследования — разработка учебного курса по английскому языку, основанного на использовании в качестве обучающих материалов фрагментов из англоязычных мультипликационных фильмов. Курс рассчитан на обучающихся, достигших уровня B1, и включает в себя несколько аспектов, в том числе фонетический.

Эффективность обучения иностранным языкам с использованием мультимедийных технологий подтверждают многие учёные (Колесникова, 2017; Лякина, Ильясова, 2018; Mayer, 2009; Stewart, 2006; Tomlinson, 2010). Подобный подход обладает такими преимуществами, как одновременное задействование двух каналов обработки информации, возможность погружения в языковую среду, интерактивность учебного процесса, повышение мотивации, возможность самостоятельной работы. Кроме того, при просмотре видеоматериалов обучающиеся могут копировать мимику и жесты говорящего, что упрощает процесс отработки произношения и помогает достичь нужного коммуникативного эффекта.

Материал и метод исследования

Материалом исследования послужили анимационные фильмы студии Disney, а также сценарии к ним. При поиске примеров, иллюстрирующих ту или иную особенность английской фонетики, применялся метод контролируемого отбора. Критериями отбора явились такие факторы, как громкость и отчётливость звука, демонстрация артикуляции, мимики и жестикуляции персонажей, простота и краткость высказывания (относительно уровня B1), значимость фрагмента с точки зрения сюжета мультфильма, его способность оказать на обучающихся эмоциональное воздействие и таким образом повысить их мотивацию.

Для разработки курса были выбраны наиболее известные мультфильмы, созданные анимационной студией Уолта Диснея (преимущественно до 1999 года). Причина такого выбора заключается в том,

¹ Научный руководитель — доцент СПбГУ, канд. филол. наук Е. А. Шамина.

что эти мультфильмы могут послужить интересным обучающим материалом как для детей и подростков, так и для взрослых, смотревших данные мультфильмы в детстве или юности. Такие мультфильмы, как «Snow White and the Seven Dwarfs», «Cinderella», «Beauty and the Beast», «Sleeping Beauty» считаются культовыми, вызывают у зрителя чувство ностальгии. Каждому знакомы главные героини этих мультфильмов — принцессы Белоснежка, Золушка, Белль и другие. Классические мультфильмы Disney славятся красочностью, замечательным музыкальным сопровождением и неповторимой сказочной атмосферой. Всё это повышает мотивацию обучающихся при прохождении курса.

Общая структура курса

По своей структуре курс состоит из разделов, каждый из которых посвящён какой-то одной теме. Каждый раздел включает в себя два проверочных теста, материалы для первичного ознакомления с темой и задания на закрепление пройденного.

Входное и итоговое тестирование

Первый тест, который также именуется входным тестированием, позволяет обучающимся определить уровень своего владения темой до начала прохождения соответствующего раздела курса. Второй тест, именуемый итоговым тестированием, находится в самом конце раздела и по характеру предлагаемых заданий идентичен первому. В случае, если задания входного тестирования покажутся ученикам слишком лёгкими, они могут перейти к следующей теме и не тратить время на просмотр всего остального раздела. В противном случае ученики смогут сохранить полученные результаты для того, чтобы отследить свой прогресс по окончании темы, поскольку чувство успеха играет немаловажную роль в поддержании мотивации.

Первичное ознакомление с новой темой

Материалы, предназначенные для первичного ознакомления с темой, включают в себя лаконичное объяснение рассматриваемого правила в виде схемы или краткого комментария и несколько иллюстрирующих его фрагментов из мультфильмов. Преимущество подобного подхода заключается в лёгкости восприятия новой темы, наглядности, практической направленности процесса обучения и возможности развития у учеников языковой интуиции.

Закрепление новой темы посредством выполнения упражнений

В числе лингводидактических методов, на применении которых основаны задания в курсе — работа с произносительной моделью, опознание, имитация, анализ тех или иных фонетических явлений, содержащихся в просматриваемых учениками видеофрагментах. Благодаря наличию видеоряда обучающиеся смогут не только слышать произношение звуков, но и видеть артикуляцию, мимику и жесты героев. Некоторые задания курса также включают в себя транскрибирование или использование изучаемой модели в контексте — например, при разыгрывании сцен из мультфильма.

Фонетические задания по теме «Слабые формы служебных слов»

Рассмотрим несколько примеров из раздела «Слабые формы служебных слов». В списке разделов курса он расположен сразу после грамматической темы «Present Perfect» и позволяет дополнительно закрепить полученные ранее знания (подобный принцип в целом, где это возможно, применяется при определении порядка тем).

Для данного фонетического раздела были отобраны видеофрагменты, содержащие такие фразы, как «*Has he kissed her yet?*», «*The young lady has disappeared*», «*Jafar has done something terrible*», «*Have we met? We have met?*», «*The Huntsman has brought me proof*», «*But don't you remember? We've met before!*». Эти фразы просты для восприятия, не слишком длинные, не перегружены трудными для произнесения словами. В то же время они связаны с ключевыми моментами из сюжета мультфильмов: речь идёт о поцелуе, способном спасти Русалочку, о таинственном исчезновении Золушки, о предполагаемой гибели Аладдина и т. д.

Задания, предлагаемые в рамках курса, также сформулированы с учётом контекста мультфильмов, что дополнительно повышает мотивацию обучающихся. Так, задание на опознание и имитацию фразы «*Jafar has done something terrible*» сформулировано следующим образом: «*What is Jasmine upset about?*». Ученику предлагается получить ответ на этот вопрос, просмотрев соответствующий видеофрагмент, и произнести услышанное вслух.

Принцип отбора фонетических тем, затронутых в курсе

Среди фонетических тем, которые затронуты в курсе — произношение гласных, слабые формы служебных слов, словесное ударение, интонация. Представляется, что эти темы требуют наибольшего внимания, потому как различия между русскими и английскими фонемами, понимание на слух и употребление в речи слабых форм служебных слов, произношение слов, содержащих более одного ударения, часто вызывает трудности у русскоговорящих.

Кроме того, поскольку разрабатываемый учебный курс ориентирован на уровень владения языком B1, в рамках исследования были проанализированы списки тем, содержащихся в наиболее авторитетных зарубежных учебных пособиях по английскому языку для уровня Intermediate (Baker, 2006; Hancock, 2003; Kelly, 2001). Темы, встречающиеся в данных пособиях наиболее часто, были отобраны для создания курса.

Результаты и выводы

Размещение подобного курса после его апробации на носителях русского языка, изучающих английский, в сети «Интернет», может расширить возможности как самостоятельного изучения американского варианта английского языка, так и овладения им под контролем преподавателя.

Список используемой литературы

1. Колесникова А. Н. Мультимедийные технологии в обучении практической фонетике. Теория и практика: пособие для студентов и аспирантов языковых вузов и учителей. М.: КДУ, Университетская книга, 2017. 126 с.
2. Лякина Е. В., Ильясова Е. В. Возможности имплементации мультимедийных технологий при обучении восприятию на слух иноязычной речи // Вестник современных исследований. 2018. № 11 (7). С. 162–165.
3. Baker A. Ship or Sheep? An Intermediate Pronunciation Course. New York: Cambridge University Press, 2006. 237 p.
4. Hancock M. English Pronunciation in Use. New York: Cambridge University Press, 2003. 201 p.
5. Kelly G. How to Teach Pronunciation. Pearson Education Limited, 2001. 164 p.
6. Mayer R. E. Multimedia Learning. 2nd ed. New York: Cambridge University Press, 2009. 304 p.

7. *Stewart D. M.* Film English: Using films to teach English // *Electronic Journal of English Education*. 2006. Vol. 24 (1). P. 1–17.
8. *Tomlinson B.* Principles and Procedures of Materials Development // *Materials in ELT: Theory and Practice*. Cambridge: Cambridge University Press, 2010. P. 81–108.

PHONETIC LYCEUM, VOL.9. ABSTRACTS (IN ENGLISH)

THE INFLUENCE OF GENDER ON THE PERCEPTION OF SPEECH RATE

Ananstasia Denisovna Ananeva
(*Saint Petersburg State University*)

Supervisor: assoc. prof., Ph.D. U.E. Kochetkova

This article explores the perception of speech tempo among Russian speakers, focusing on the influence of gender on the perception of both speech rate and articulation rate. Using the recordings of reading from the corpus CoRuss, pairs of male and female speakers were selected based on similar speech rates. Participants listened to recordings and determined which speaker sounded faster. Results indicate that external factors such as pauses filled with extraneous noise significantly influenced male participants' responses, suggesting potential gender differences in speech perception. While gender of the listener did not affect neither full rate perception nor articulation rate perception, the gender of the speaker did impact the perception of articulation rate. Male speakers were more often perceived as speaking faster, aligning with previous research suggesting gender biases in speech perception. This study underscores the multifaceted nature of tempo perception, influenced not only by linguistic factors but also by social variables such as speaker gender.

PERCEPTION OF INTONATION IN SYNTHESIZED SPEECH

Daria Sergeevna Gineva
(*Saint Petersburg State University*)

Supervisor: assoc. prof., Ph.D. U.E. Kochetkova

This study investigates the prosodic characteristics and naturalness of synthesized speech, focusing on gender and age variations. Analyzing synthesized speech from prominent synthesizers, including "Sber Salut Speech", "Zvukogram" and "Yandex SpeechKit", significant differences in melodic contours between male and female voices were observed. While male voices generally exhibited a narrower melodic range, they were perceived as more natural in meaningful contexts. However, this perception shifted with meaningless ("nonsense") utterances, where female voices received higher naturalness ratings. Furthermore, elderly voices were consistently rated as more natural than younger ones.

AUTOMATIC DETECTION AND NEUTRALIZATION OF EMOTIONAL SPEECH

Ivan Evgenievich Gurkov

(Saint Petersburg State University)

Supervisor: assoc. prof., Ph.D. O.A.Mitrofanova

This work explores automatic detection and neutralization of emotional speech using modern natural language processing techniques like sentiment analysis and style transfer. The dataset consists of transcribed emergency call recordings labeled with emotional categories. Machine learning models trained on linguistic features and acoustic features achieved around 62% accuracy in emotion classification. To neutralize emotional utterances while preserving semantics, a sequence-to-sequence model was fine-tuned on data generated by prompting a large language model (GPT-4) to rephrase the emotional texts neutrally. The fine-tuned model attained a ParaScore of 0.826, demonstrating its capability for emotion neutralization while retaining meaning. The results showcase the potential of leveraging sentiment analysis and style transfer to improve digital communication quality.

WRITTEN AND ORAL TEXT THROUGH THE PRISM OF KEYPHRASE EXTRACTION

Darya Dmitrievna Guseva

(Saint Petersburg State University)

Supervisor: assoc. prof., Ph.D. O.A.Mitrofanova

The article presents the findings of a study on the perception of key phrases in written and oral texts. The research was conducted on the basis of audio recordings of lectures from the project «ПостНаука». The study analyzes differences in the way texts are perceived by individuals from different age groups. Based on the analysis, it can be concluded that there are variations in the way different age groups perceive written and oral texts. Adults tend to be more consistent in selecting keyphrases compared to teenagers. This could indicate that adults may be more selective in their choice of keyphrases and attach greater significance to certain ideas within the text.

EMOTIONAL SPEECH RECOGNITION OF SEVERE SHOCK SURVIVORS

Mikhail Dolgushin

(Saint Petersburg State University)

Supervisor: assoc. prof., Ph.D. V.V.Evdokimova

Recognition of highly emotional speech remains a poorly studied part of the field of speech technologies. The purpose of this work is to conduct experiments on

highly emotional speech recognition based on interviews provided by the Yad Vashem Foundation. A system is created capable of recognizing speech based on neural network models using chosen material. The highest WER of the resulting system is 35,88 % on recordings with one speaker and 39,89 % on audio with two speakers. An analysis of discrepancies between automated transcriptions and human transcriptions is performed. The presented results contribute to the research field of emotional speech recognition.

PROSODIC CORRELATES OF POLITENESS IN ORAL RUSSIAN SPEECH. PILOT EXPERIMENT

Victoria Alexandrovna Lozhkina
(Saint Petersburg State University)

Supervisor: assoc. prof., Ph.D. U.E. Kochetkova

This report presents the results of a pilot study aimed at identifying the key acoustic features of polite and impolite speech. The increasing development of communication opportunities and channels, as well as the emergence of more advanced artificial voice assistant systems, has led to the need to address the issue of prosodic features that organize polite speech. The material for the analysis was the results of a pilot perceptual experiment conducted in December 2023. Native speakers of the Russian language were invited to listen to isolated audio recordings (lasting from 2 to 11 seconds) in different communicative situations (a public interview, dialogue) and with different settings for the speaker. They were asked to determine how much each recording sounded polite and empathetic. They also had to indicate whether they wanted to continue communicating with the speaker and whether the social distance between them was great. The recordings were phrases with and without a lexical marker of politeness (“Please tell us what kind of award it is” and “The stop is right around the corner”). The analysis was performed manually using the Praat and Wave Assistant programs. As a result, a correlation has been found between a decrease in the level of politeness in speech and the severity of changes in the pitch, and, more generally, the “fragmentation” of intonation. This can be seen in the example of strong accentuation on the last syllable of a phrase. In addition, conclusions have been drawn about the overall pace of speech, which is influenced by factors such as the pragmatics of the message, the distance between speakers, and the desire to continue the conversation. Finally, there is a correlation between an increase in the overall tone of speech in impolite situations compared to polite ones, as spoken by the same person. These findings provide a basis for further research on the prosodic aspects of politeness by the author.

PHONETIC CHARACTERISTICS OF SPEECH UNDER COGNITIVE LOAD

Mariia Romanovna Maksimova
(*Saint Petersburg State University*)

Supervisor: assoc. prof., Ph.D. V. V. Evdokimova

This paper analyses phonetic features of speech under cognitive load, without cognitive load and in final lengthening. Previous research revealed that cognitive load is associated with increasing fundamental frequency (F0), increasing intensity, laryngealization, narrowing F0 range, changing speech tempo and articulation rate. To study speech under cognitive load, Stroop test was used. Seven female SPbU students participated. They were presented words denoting colour names. Each word was printed in a colour that did not match the word meaning. The main task was to pronounce the font colour. As a dual task, subjects answered questions related to their field of studies. In order to analyze speech without significant cognitive load, participants read a phonetically representative text. Reading samples were also used to discover the acoustic parameters of speech sounds in final lengthening. The findings showed that the duration of both vowels and consonants increased under cognitive load compared to the absence of cognitive load and to final lengthening. This implies that articulation rate lowers under cognitive load. F0 range was larger under cognitive load than under the other conditions. The duration of laryngealized segments was the greatest under cognitive load.

PHONETIC INTERFERENCE IN RUSSIAN-L2 KOREAN SPEAKERS

Marina Alekseevna Pakhomova
(*Saint Petersburg State University*)

Supervisor: assoc. prof., Ph.D. S. O. Tananaiko

This paper examines phonetic interference experienced by native Korean speakers speaking Russian. The objective of this research is to identify the characteristics of Russian phonetics which pose the greatest difficulties for Korean native speakers. Using both qualitative and quantitative methods, the paper utilises recordings of Korean native speakers speaking Russian as the basis for the analysis. This paper identified several consistent mistakes engendered by Korean phonetics that Korean native speakers made while speaking Russian. These included voiceless pronunciation of voiced consonants, affricatization of the Russian /ž, z, z', t', d'/, replacement of /č/ to /tɕ/ and of /i/ to /i/ or /u/. The findings of this paper contribute to the development of rules and principles for teaching Russian phonetics to Korean native speakers and identifying particular sounds that Korean native speakers find challenging.

RUSSIAN-CHINESE INTERFERENCE IN THE
IMPLEMENTATION OF INTONATION IN SPEECH IN THE
RUSSIAN LANGUAGE

*Piskunova Zhanna Evgenevna,
Saint-Petersburg Peter the Great Polytechnic University
Supervisor: assoc. prof., Ph.D. M.P. Agafonova*

The article “Russian-Chinese interference in the implementation of intonation in speech in the Russian language” will present a study aimed at identifying specific problems that Chinese speakers of the Russian language have in the process of learning and using the correct intonation in Russian speech. The article discusses in detail the features of Russian intonation and its influence on the meaning of the statement, as well as analyzes the ways of its implementation and transition from one intonation to another. The article discusses various causes of interference in intonation, for example, the features of the sound structure of speech and its rhythmic and intonation features, which differ from the Russian language. The author of the article explores the specific problems associated with the transfer of intonation from Chinese to Russian, and suggests methods that can help Chinese speakers of the Russian language to correctly implement intonation and, therefore, convey the meaning of the statement. The results of the study are especially important for teachers of the Russian language who teach Chinese speakers, as well as for everyone who is interested in the problems of interlingual communication and intercultural communication in education. The article also contains methodological materials that can be used in studying the correct implementation of intonation in Russian speech for Chinese students, helping them to successfully master this difficult part of the language being studied.

AUDIOVISUAL DESIGN OF THE PRAGMATIC MARKER V
OBSHCHEM (ANYWAY)

*Evgenia Alekseevna Radionova
(Saint Petersburg State University)
Supervisor: assoc. prof., Ph.D. U.E. Kochetkova*

The study is devoted to the pragmatic marker (PM) *v obshchem* (anyway), namely, how its intonation design correlates with the gesture that the speaker uses when pronouncing the unit under study. The report examines the contexts with PM, taken from the multimedia subcorpus of the National Corpus of the Russian Language (NKRYA). The material was placed using the Wave Assistant and ELAN programs. After processing the material, the results of the analysis of the intonation design of the PM *v obshchem* (anyway) and the analysis of the gestures used by the speakers when pronouncing the unit in question were compared. As a result of the study, it was revealed that the pragmatic marker *v obshchem* (anyway), as a rule, is accompanied by a gesture, and also in most cases the upward

movement of the tone in this unit correlates with the fact that speakers, using this marker, make an upward movement with their heads, shoulders, arms, and eyes.

AUTOMATIC TRANSCRIPTION OF RUSSIAN SPEECH USING KALDI TOOL

Lada Leonidovna Skorobagatko
(Saint Petersburg State University)

Supervisor: assoc. prof., Ph.D. E.I. Riekhakaynen

The study is focused on automating speech transcription. The tool Kaldi is used to train models on the data from the Russian Spoken Corpus. 31 models were trained, the model with words as dictionary entries showed the worst result, while the algorithm using allophone dictionary demonstrated the best Word Error Rate metric. Changing the parameters of the model (in particular, the length of N-grams) also improved the results: it was possible to reduce the Word Error Rate by 4 %.

ERRORS IN READING HOMOGRAPHS AND SOME OTHER FACTORS IN THE PERCEPTION OF SYNTHESIZED SPEECH

Ivan Igorevich Stoliarov
(Saint Petersburg State University)

Supervisor: assoc. prof., Dr. O.A. Mitrofanova

The study focuses on the role of correct or incorrect reading of homographs in assessing the quality of the latest generation of Text-to-Speech systems. The study involves an experiment on the perception of natural and synthesized speech. Natural speech samples were taken from the fully annotated corpus of Russian speech CORPRES; synthesized speech samples were generated in demo versions of the SaluteSpeech and Yandex SpeechKit TTS systems. The results of the experiment demonstrated the importance of correct reading of homographs (even when achieving high quality of other speech parameters), and also allowed us to formulate observations about some other factors affecting the perception of synthesized speech.

PROSODIC FEATURES OF THE BEGINNING OF A LECTURE AND THEIR IMPACT ON THE AUDIENCE

Titova Renata Ruslanovna
(Saint Petersburg State University)

Supervisor: assoc. prof., Ph.D. U.E. Kochetkova

This paper presents the results of an analysis of the prosodic features of the speech used by two teachers at the start of their lectures, as well as the results of a trial experiment involving 46 female and 10 male respondents. The respondents were asked to rate their perception of the sound of the lecturers' voices, their level of

interest in the material, and the emotions they felt while listening to the lectures. The analysis revealed correlations between the prosodic characteristics of the lecturers' speech (such as tempo, intonation patterns, types of pauses, and accents) and the reactions of the audience. The study found that certain prosodic features can influence listeners' perception of the speaker's personality, level of interest, and overall effectiveness as a communicator.

USING ANIMATED FILMS AT ENGLISH CLASSES

Nadezhda Aleksandrovna Filimonova

(Saint Petersburg State University)

Supervisor: assoc. prof., Ph.D. E. A. Shamina

The goal of the study is to create an English course based on authentic animated films. Among the phenomena that are going to be taught as a part of the phonetical aspect of the course are vocals, weak forms of function words and secondary stress in words. The effectiveness of teaching foreign languages using multimedia technologies has been confirmed by many scientists. This approach has such advantages as the simultaneous use of two channels of information processing, the possibility of immersion in the language environment, the interactivity of the educational process, increased motivation, and the possibility of independent work. In addition, when watching video materials, students can copy the speaker's facial expressions and gestures, which simplifies the process of practicing pronunciation and helps achieve the desired communicative effect.