ЭКСПЕРИМЕНТАЛЬНАЯ ОЦЕНКА ВКЛАДА ЭЛЕМЕНТОВ КОМПИЛЯЦИИ В ПРАВДОПОДОБИЕ СИНТЕЗИРОВАННОГО РЕЧЕВОГО КЛОНА[1]

Л.И. Цирульник *(liliya_tsirulnik@ssrlab.com)*
Б.М. Лобанов *(lobanov@newman.bas-net.by)*
Объединённый институт проблем информатики НАН Беларуси, Минск, Беларусь

Работа выполнена в рамках продолжающихся исследований по клонированию голоса и дикции личности. Описываются результаты эксперимента по оценке влияния элементов компиляции различного фонетического типа (ударные и безударные гласные, согласные) и различного уровня (аллофоны и мультифоны) на восприятие персональных фонетико-акустических характеристик голоса и дикции при синтезе речи по тексту. В экспериментах используется общепринятая методика субъективной оценки качества синтезированной речи, так называемая MOS-оценка. Обсуждаются полученные результаты и перспективы использования различных по уровню элементов компиляции в системах создания речевых клонов личности.

Введение

Задача клонирования голоса и дикции личности, т.е. максимального приближения характеристик синтезированной речи к персональным характеристикам естественной речи конкретного диктора, была поставлена и затем детализирована в [1, 2]. При этом основными характеристиками речи диктора считаются акустические свойства голоса (тембр, высота голоса и др.), фонетические особенности произношения (акцент, особенности дикции и др.) и просодические (интонационные, динамические, ритмические) свойства речи диктора.

При персонализированном синтезе речи по тексту используется компиляционный метод [2] как наиболее подходящий для воспроизведения индивидуальных особенностей речи диктора [3]. Компиляционный синтезатор осуществляет преобразование входного орфографического текста в звучащую речь, при этом используются как общеязыковые правила, так и индивидуальные особенности генерации речи конкретным диктором. В частности, для воспроизведения при синтезе индивидуальных фонетических и акустических особенностей речи используются элементы компиляции, являющиеся отрезками естественной речевой волны конкретного диктора.

К настоящему времени «ручным» способом [4] и с помощью автоматизированной системы клонирования [5,6] получены клоны нескольких мужских и женских голосов. В данной работе ставятся и решаются следующие задачи:

- разработать адекватную методику оценки степени сходства синтезированного клона с естественной речью (т. е. правдоподобие речевого клона),
- получить численную оценку правдоподобия наилучшего из созданных речевых клонов,
- оценить вклад элементов компиляции различного уровня (аллофоны, диаллофоны) в правдоподобие речевого клона,
- оценить вклад фонем того или иного типа (ударные и безударные гласные, согласные) в правдоподобие речевого клона.

1. Методика оценки правдоподобия речевого клона

Существует несколько методов оценки качества синтезированной речи [7-10], основанной на расчёте корреляции между естественным и синтезированным речевыми сигналами в пространстве различных параметров сигнала. Однако даже лучшие из них не дают результат, приближающийся к результатам субъективной оценки. Поэтому в экспериментах по определению степени сходства синтезированного клона с естественной речью (т. е. правдоподобия речевого клона) предпочтение было отдано оценке субъективного мнения, так называемой MOS-оценке. Методика проведения эксперимента основывалась на Рекомендации P.85 ITU-T «Метод субъективной оценки качества речи устройств речевого вывода» [11], но была адаптирована для данной задачи.В связи с тем, что оценивалось не качество синтезируемой речи, а правдоподобие речевого клона, были подходящим образом скорректированы форматы стимулов, опросные листы и процедура прослушивания.

1.1. Подготовка фонетико-акустических баз

Для экспериментов использованы записи естественного голоса диктора БЛ и его синтезированного клона, а также клона голоса диктора АТ. Оба выбранных для эксперимента голоса имели примерно одинаковый диапазон изменения частоты основного тона (высоту голоса). Клоны голосов двух дикторов – БЛ и АТ –синтезированы с

использованием соответствующих фонетико-акустических БД. Каждая из БД содержит полный набор звуковых волн аллофонов. В состав БД диктора БЛ мог быть включён, кроме того, набор звуковых волн диаллофонов, в который вошли как наиболее трудные для «вычленения» из речевого потока сочетания аллофонов (например, гласный – гласный, гласный и “J”), так и наиболее часто встречающиеся в речи сочетания (например, “NA”, “PA”, “PO” и др.). Исходным материалом для подготовки фонетико-акустических баз послужили записи естественной речи, выполненные в студийных условиях.

На основе двух фонетико-акустических БД клонов сформированы 7 фонетико-акустических баз, содержимое которых описано в таблице 1.

Название базы	**Содержимое базы**
База 1	Полный набор аллофонов и диаллофонов диктора БЛ
База 2	Полный набор аллофонов диктора БЛ (без диаллофонов)
База 3	Аллофоны гласных диктора БЛ+ аллофоны согласных диктора АТ
База 4	Аллофоны ударных гласных диктора БЛ+ аллофоны безударных гласных и согласных диктора АТ
База 5	Аллофоны безударных гласных и согласных диктора БЛ+ аллофоны ударных гласных диктора АТ
База 6	Аллофоны согласных диктора БЛ+ аллофоны гласных диктора АТ
База 7	Полный набор аллофонов диктора АТ

Таблица 1. Содержимое фонетико-акустических баз для тестирования

Как следует из таблицы 1, для экспериментов были сформированы две БД для синтеза клонов диктора БЛ (базы 1, 2) и одна БД для синтеза клонов диктора АТ (база 7). Кроме того, сформированы четыре БД (базы 3 – 6), с помощью которых синтезировались «клоны-химеры», обладающие в той или иной степени свойствами голоса дикторов БЛ или АТ.

Все элементы баз сохранялись в формате WAVE PCM с частотой дискретизации 22050Гц и разрядностью 16 бит.

1.2. Подготовка стимулов для тестирования

В качестве сообщений были подобраны 20 фонетически сбалансированных фраз. Каждая фраза состояла из последовательности трёх-четырёх слов.

Все фразы были произнесены диктором БЛ в студийных условиях, идентичных условиям для подготовки фонетико-акустических баз клона. Запись производилась на цифровые носители, и была сохранена в формате WAVE PCM с частотой дискретизации 22050 Гц и разрядностью 16 бит. Фразы были произнесены со средним темпом речи, с небольшими вариациями частоты основного тона, с интонацией перечисления. Длительность пауз между словами во фразе была приведена к значению 200 мс. Длительность каждой из фраз составила 3,2 - 4,3 секунды. Диапазон частоты основного тона составил 80-120 Гц.

На основании подготовленных фонетико-акустических баз были синтезированы 7 групп фраз-клонов. Каждая группа состояла из 20 фраз, идентичных по содержанию фразам, произнесённым естественным голосом. Для устранения влияния просодических характеристик на восприятие синтезированной речи во всех синтезированных фразах сохранялись примерно тот же темп речи, длительность пауз между словами, амплитуда сигнала и значения частоты основного тона, характерные для естественных фраз.

Стимулами для тестирования и оценки являлись пары фраз одинакового содержания. При этом первая фраза в паре являлась записью естественной речи диктора БЛ, а вторая фраза могла быть либо записью речи того же диктора с искусственно внесёнными незначительными мультипликативными искажениями, либо записью его синтезированного клона, полученного в соответствии с таблицей 1.

Таким образом, было сформировано 20*8=160 стимулов, которые представлялись аудиторам в случайном порядке.

Пауза между фразами в паре составляла 700 мс, пауза между стимулами – 5 секунд.

1.3. Проведение эксперимента

Для оценки правдоподобия синтезированного речевого клона аудиторам было предложено ответить на вопрос «Похож ли второй из услышанных голосов на первый?», используя шкалу оценки, представленную в таблице 2.

Каждый стимул аудиторы прослушивали один раз. Для того, чтобы аудиторы сфокусировали внимание на сходстве голосов, а не на разборчивости произносимых фраз (для минимизации напряжения при прослушивании) текстовое содержание каждого из стимулов было записано на опросном листе.

Оценка	*Значение*

1	Нет, совсем другой голос
2	Нет, пожалуй, это другой голос
3	Немного похож
4	Да, очень похож
5	Да, практически тот же голос

Таблица 2. Шкала оценки сходства голосов

Аудиторами являлись 8 мужчин в возрасте от 21 до 60 лет, носители русского языка, без выявленных дефектов слуха. Тест проходил в тихой комнате, длился около 50 минут и был разбит на 2 сессии по 25 минут.

2. Обработка результатов эксперимента

Обобщенная оценка правдоподобия речевых клонов (MOS-оценка) выражается через среднее значение и дисперсию оценок всех аудиторов в соответствии со шкалой сходства голосов (таблица 2) по каждому из тестируемых типов БД (таблица 1).

Для определения статистической значимости оценок, полученных для различных типов голосов, был осуществлён однофакторный дисперсионной анализ результатов с использованием F-критерия и множественное попарное сравнение [12] с использованием критерия Тьюки достоверно значимой разности.

При анализе результатов вычислялась дисперсия ошибки $SS_{ошибки}$ и дисперсия эффекта $SS_{эффекта}$ в соответствии с формулами (1)-(3):

$$SS_{(общая)} = \left(\sum x_1^2 + \sum x_2^2 + \ldots + \sum x_r^2\right) - \left(\sum x_1 + \sum x_2 + \ldots + \sum x_r\right)/N \quad (1),$$

$$SS_{ошибки} = \frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \ldots + \frac{(\sum x_r)^2}{n_r} - (\sum x_1 + \sum x_2 + \ldots + \sum x_r)/N \quad (2),$$

$$SS_{эффекта} = SS_{(общая)} - SS_{ошибки} \quad (3),$$

где N – общее количество наблюдений (в данном тесте равное 1280),

x_i –значение наблюдения в i-той группе,

n_i – количество наблюдений в группе i (в данном тесте равное 160 для каждой группы),

r – количество групп (в данном тесте равное 8).

Затем вычислялось количество степеней свободы эффекта $DF_{эффекта}$ и ошибки $DF_{ошибки}$ (формулы (4), (5)) и, на их основе, значения среднего квадрата эффекта $MS_{эффекта}$ и среднего квадрата ошибки $MS_{ошибки}$ (формулы (6), (7)):

$$DF_{эффекта} = r - 1 \quad (4),$$

$$DF_{ошибки} = N - r \quad (5),$$

$$MS_{эффекта} = SS_{эффекта} / DF_{эффекта} \quad (6),$$

$$MS_{ошибки} = SS_{ошибки} / DF_{ошибки} \quad (7)$$

Пропорция средних квадратов F, необходимая для вычисления статистической значимости, определялась в соответствии с формулой (8):

$$F = MS_{ошибки} / MS_{эффекта} \quad (8)$$

Результаты вычислений представлены в таблице 3.

	SS	MDF	MS	F	p
Эффект	2034.3	N7	290.6	425.61	<0.0001
Ошибка	868.6	O1272	0.7		

Таблица 3.Результаты однофакторного дисперсионного анализа

Для множественного сравнения для каждой пары групп ***i*** и ***j*** было вычислено значение разности $MS_{попарное}$:

(9),

$$MS_{\text{попарное}} = (M_i - M_j) / \sqrt{MS_{\text{ошибки}} \left(\frac{1}{n}\right)}$$

где M_i, M_j – средние значения оценок для групп i,j соответственно,

n – количество наблюдений в группе.

Затем были вычислены границы разности с доверительным интервалом 95% и сделан вывод о значимости/не значимости разности.

Результаты вычислений представлены в таблице 4.

Пары групп	$MS_{\text{попарное}}$	95%-ный доверительный интервал		Значимость разности
		Нижняя граница	Верхняя граница	
Естественный – База1	0.5	0.3	0.8	Значима
Естественный – База2	0.8	0.6	1.1	Значима
Естественный – База3	1.3	1.0	1.5	Значима
Естественный – База4	2.7	2.4	2.9	Значима
Естественный – База5	2.8	2.6	3.0	Значима
Естественный – База6	3.4	3.2	3.6	Значима
Естественный – База7	3.3	3.1	3.6	Значима
База1 – База2	0.3	0.5	1.0	Значима
База1 – База3	0.8	0.2	0.7	Значима
База1 – База4	2.2	1.9	2.4	Значима
База1 – База5	2.3	2.0	2.5	Значима
База1 – База6	2.9	2.6	3.1	Значима
База1 – База7	2.8	2.6	3.1	Значима
База2 – База3	0.5	0.2	0.7	Значима
База2 – База4	1.8	1.6	2.1	Значима
База2 – База5	2.0	1.7	2.2	Значима
База2 – База6	2.6	2.3	2.8	Значима
База2 – База7	2.5	2.3	2.7	Значима
База3 – База4	1.4	1.1	1.6	Значима
База3 – База5	1.5	1.3	1.8	Значима
База3 – База6	2.1	1.9	2.4	Значима
База3 – База7	2.0	1.8	2.3	Значима
База4 – База5	0.1	-0.1	0.4	Не значима
База4 – База6	0.7	0.5	1.0	Значима
База4 – База7	0.7	0.4	0.9	Значима
База5 – База6	0.6	0.3	0.9	Значима
База5 – База7	0.5	0.3	0.8	Значима
База6 – База7	-0.1	-0.3	0.2	Не значима

Таблица 4.Результаты множественного сравнения с использованием критерия Тьюки

3. Обсуждение результатов

Наилучшую численную оценку правдоподобия среди созданных речевых клонов получил клон диктора БЛ (база 1), в котором использована БД звуковых волн аллофонов и диаллофонов в количестве 1852 единиц. На рис.1 приведены MOS-оценки правдоподобия речевого клона диктора БЛ, полученного с использованием базы 1, клона диктора АТ, полученного с использованием базы 7, а также (для сравнения) MOS-оценка правдоподобия естественного речевого сигнала диктора БЛ с искусственно внесёнными незначительными мультипликативными искажениями. Слева на рисунке представлены значения шкалы оценок, а для каждого типа голоса показано количественное значение средней оценки и дисперсия.

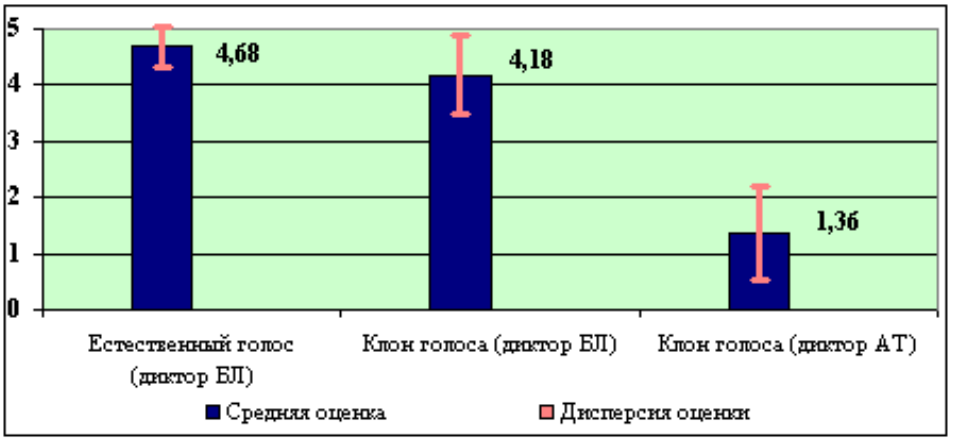


Рис. 1. MOS-оценка правдоподобия речевых клонов двух дикторов в сравнении с оценкой естественного голоса диктора БЛ

Как видно из рис. 1, достигнутая оценка правдоподобия речевого клона диктора БЛ – 4,18 при дисперсии 0,7 – близка к оценке естественной речи и существенно отличается от оценки, полученной для клона диктора АТ. Как следует из таблицы 3, полученные результаты являются статистически значимыми.

Оценка вклада в правдоподобие речевого клона элементов компиляции различного уровня (аллофонов и диаллофонов) иллюстрируется рис. 2, где приведены MOS-оценки правдоподобия клонов диктора БЛ, полученные с использованием базы 1 (полный набор аллофонов и диаллофонов) и базы 2 (набор аллофонов диктора БЛ без диаллофонов), а также (для сравнения) MOS-оценка правдоподобия естественного речевого сигнала диктора БЛ с искусственно внесёнными незначительными мультипликативными искажениями.

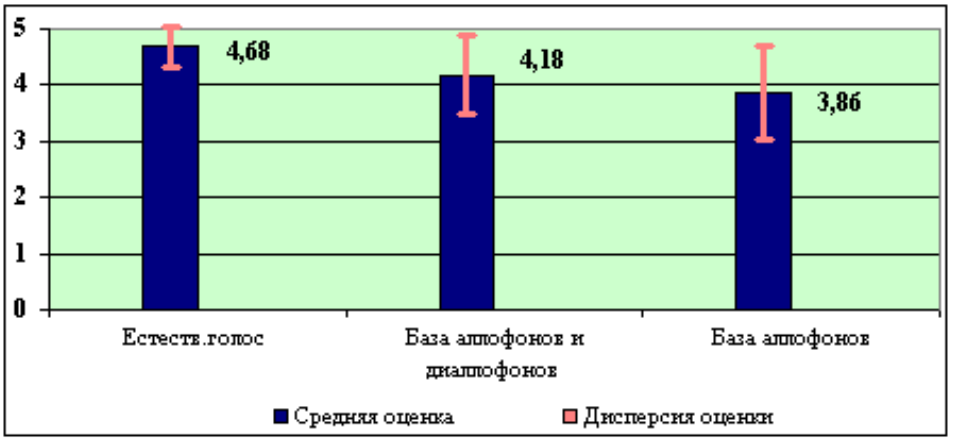


Рис. 2. MOS-оценка правдоподобия речевых клонов диктора БЛ на основе двух различных баз в сравнении с оценкой естественного голоса диктора БЛ

Как видно из рис. 2, добавление в БД диаллофонов даёт ощутимый эффект в восприятии правдоподобия речевого клона, причём, как следует из таблицы 4, разница в оценках базы 1 и базы 2 является статистически значимой.

Оценка вклада в правдоподобие речевого клона фонем того или иного типа (ударные и безударные гласные, согласные) иллюстрируется на рис. 3. Здесь приведены MOS-оценки правдоподобия клонов дикторов БЛ и АТ, полученные с использованием, соответственно, базы 2 (набор аллофонов диктора БЛ) и базы 7 (набор аллофонов диктора АТ). Кроме того, приведены MOS-оценки для четырёх БД (базы 3 – 6), с помощью которых синтезировались «клоны-химеры», обладающие в той или иной степени свойствами голоса дикторов БЛ или АТ, а именно:

- база 3, в которой использованы аллофоны всех гласных диктора БЛ, а аллофоны согласных взяты от диктора АТ,
- база 4, в которой использованы аллофоны только ударных гласных диктора БЛ, а аллофоны безударных гласных и согласных взяты от диктора АТ,
- база 5, в которой использованы аллофоны безударных гласных и согласных диктора БЛ, а от диктора АТ – только аллофоны ударных гласных,
- база 6, в которой использованы только аллофоны согласных диктора БЛ, а от диктора АТ –аллофоны всех гласных.

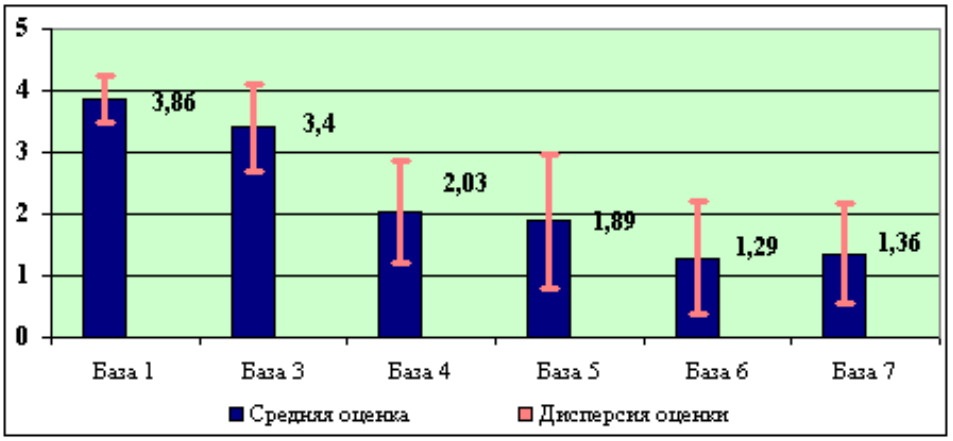
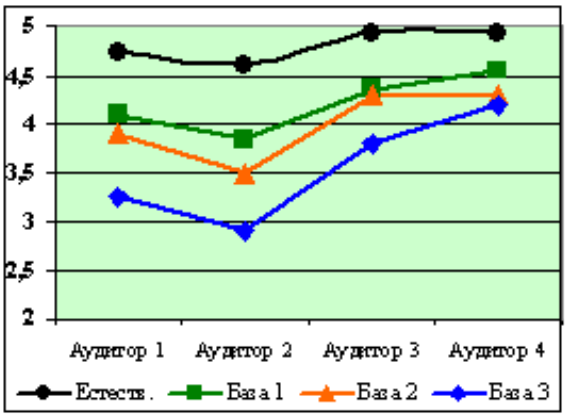


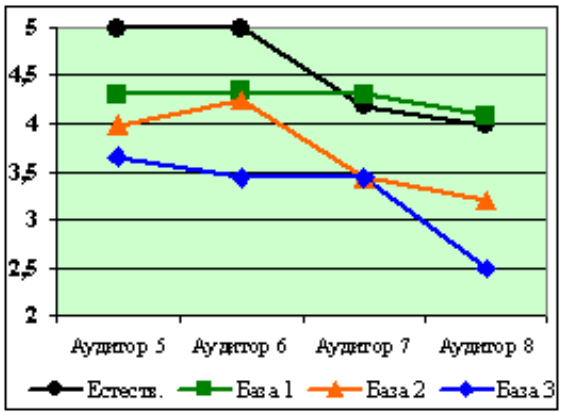
Рис. 3.MOS-оценка правдоподобия речевых клонов на основе семи различных баз

Как видно из рис. 3, наибольший вклад, как и ожидалось, в правдоподобие клона вносит комплекс ударных и безударных гласных (база 3). Использование в клоне только ударных гласных (база 4) или только безударных гласных и согласных (база 5), хотя и повышает правдоподобие клона БЛ в сравнении с клоном АТ, однако не столь значительно. Кроме того, как следует из таблицы 4, разница между оценками базы 4 и базы 5 не является статистически значимой. Можно утверждать, что отсутствие в клоне либо ударных гласных, либо безударных гласных и согласных клонируемого диктора одинаково ощутимо уменьшает правдоподобие создаваемого речевого клона. Замена в базе клона АТ только согласных на соответствующие согласные из базы клона БЛ не приводит к сколь-нибудь существенному изменению правдоподобия клона.

Интересно отметить, что персональное ранжирование голосов по степени сходства не всегда совпадает со средним. Персональные MOS-оценки каждым из аудиторов правдоподобия речевых клонов для баз 1, 2, 3, а также (для сравнения) правдоподобия естественного речевого сигнала диктора БЛ показаны на рис. 4 (а,б). Значения шкалы оценок показаны в диапазоне 2 – 5.



а)



б)

Рис. 4. MOS-оценки правдоподобия речевых клонов а) аудиторами 1-4, б) аудиторами 5-8

Заслуживает внимания тот факт, что два из восьми аудиторов (аудиторы 7 и 8) нашли речевой клон, синтезированный с использованием базы 1 (набор аллофонов и диаллофонов), одинаково или даже более «правдоподобным», чем естественный искажённый сигнал. В то же время три из восьми аудиторов (аудиторы 1,3,6) нашли почти одинаково правдоподобными речевые клоны, синтезированные с использованием базы 1 (содержащей и аллофоны, и диаллофоны) и базы 2, содержащей только аллофоны. Этот факт может быть объяснён тем, что содержащиеся в базе 1 диаллофоны использовались не в каждой из синтезируемых фраз. Действительно, после того, как было выделено подмножество фраз, при синтезе которых использовались диаллофоны, результаты показали, что

вычисленная на выбранном подмножестве MOS-оценка правдоподобия клона, полученного с использованием базы 1, превышает MOS-оценку правдоподобия клона, полученного с использованием базы 2, для каждого аудитора.

Выводы

Проведенная экспериментальная оценка правдоподобия синтезированного речевого клона показала, что разработанная методика клонирования персональных фонетико-акустических особенностей речи диктора и используемая технология автоматизированного создания фонетико-акустических речевых БД, содержимое которых используется в качестве элементов компиляции при синтезе речи, позволяет создавать синтезированные речевые клоны с достаточно высокой степенью правдоподобия. Полученный результат позволяет надеяться, что дальнейшее повышение качества клонирования может быть достигнуто путём наращивания БД диаллофонами, а также мультифонами более высокого уровня. Направления дальнейших исследований будут связаны также с автоматизацией процесса анализа просодических особенностей речи диктора и создания персональных просодических БД.

Список литературы

1. *Лобанов Б.М. и др*. Синтезатор речи по тексту как компьютерное средство “клонирования” персонального голоса. // Международная конференция «Диалог-2001». Сб. науч. тр. М., 2001. С 265-272.
2. *Лобанов Б. М.* Компьютерное «клонирование» персонального голоса и речи // Новости искусственного интеллекта. №5(55). М., 2002. С. 35-39.
3. *Лобанов Б.М*. Синтез речи по тексту // Четвёртая Международная летняя школа-семинар по искусственному интеллекту. Сб. науч. тр. Мн.:Изд. БГУ, 2000. С. 57-76.
4. *Lobanov B.M, Karnevskaya E.B*. TTS-Synthesizer as a Computer Means for Personal Voice Cloning (On the example of Russian) // Phonetics and its Applications. Stuttgart: Franz Steiner Verlag, 2002.P. 445-452.
5. *Лобанов Б.М., Киселёв В.В*. Автоматизация клонирования персонального голоса и дикции для систем синтеза речи по тексту // Международная конференция «Диалог-2003».Сб. науч. тр. М, 2003. С. 417-424.
6. *Цирульник Л.И.* Автоматизированная система клонирования фонетико-акустических характеристик речи // Информатика. № 1(9).Мн., 2006. С. 37-46.
7. *Thorpe L., Yang W.* Performance of current perceptual objective speech quality measures. // Proceeding of IEEE Workshop on speech coding. 1999. P. 144-146.
8. *Chen J.-D., Campbell N*. Objective distance measures for assessing concatenative speech synthesis. // Proceedings of EuroSpeech’1999. 1999. V. 2. P. 611-614.
9. *Chu M., Peng H.* An objective measure for estimating MOS of synthesized speech. // Proceedings of EuroSpeech’2001. 2001. P.2087-2090.
10. *Wouters J., Magon M.* A Perseptual evaluation of Distance Measures for Concatenative Speech Synthesis. // Proceedings of ICSPL’98. 1998. P.2747-2750.
11. A method for subjective performance assessment of the quality of speech voice output devices. ITU-T Recommendation P.85. ITU-T.1994.
12. *Кендал М. Дж., Стюарт А*. Статистические выводы и связи.М.:Наука, 1973.

[1] [1]*Работа выполнена при поддержке европейского фонда INTAS в рамках проекта «Разработка многоголосовой и многоязыковой системы синтеза и распознавания речи (языки: белорусский, польский, русский)» в соответствии с грантом INTAS № 04-77-7404.*