АЛГОРИТМ ТЕКСТОНЕЗАВИСИМОГО ОБУЧЕНИЯ ДЛЯ СИСТЕМ МУЛЬТИГОЛОСОВОГО СИНТЕЗА РЕЧИ

В. А. Захарьев, А. А. Петровский
Кафедра систем управления
Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: {zahariev, palex}@bsuir.by

В докладе рассмотрены вопросы применения технологии конверсии голоса для построения мультиголосовых систем синтеза речи по тексту (МГСРТ). Представлен разработанный метод и алгоритм текстнезависимого обучения для МГСРТ, позволяющий проводить обучение системы на основе непараллельных корпусов обучающих данных.

Введение

Система мультиголосового синтеза речи - эта система синтеза речи по тексту (СРТ), которая позволяет вести синтез голосом произвольного (целевого) диктора. Актуальность создания систем данного рода обуславливается устойчивой тенденцией к персонализации устройств и программного обеспечения наблюдающейся на рынке речевых технологий. Для функционирования система должна быть предварительно настроена на голос целевого диктора. Стандартные методы добавления голосов основаны на создании акустических баз для каждого диктора являются весьма затратными с точки зрения стоимости и скорости разработки. Для решения данной проблемы предлагается применить конверсию голоса, которая является техникой обработки речевого сигнала, позволяющей реализовать процесс трансформации параметров голоса, характеризующих речь исходного диктора (ИД), в параметры целевого (ЦД)[1].

I. Метод текстонезависимого обучения

Для построения функции конверсии ИД в ЦД необходимо проведение этапа обучения, который включает в себя стадию сопоставления векторов параметров, характеризующих голоса дикторов, полученных в результате анализа и параметризации сигнала. Обучение может осуществляться на основе тектозависимого и текстонезависимого подходов. При текстозависимом обучении ИД и ЦД обязаны начитывать один и тот же текст. Общий текст подразумевает общее фонетическое составляющее, варьирующееся в определенных пределах (в зависимости от пола, возраста, манеры и др. особенностей диктора), однако, принципиально одинаковое с точки зрения фонетики. Это дает возможность провести временное сопоставление векторов параметров, и в дальнейшем сопоставить кластеры акустических подпространств параметров между собой. Недостатком такого подхода являются трудозатраты на подготовку таких параллельных корпусов. В случае же текстонезависимого обучения дикторы вольны начитывать в микрофон произвольный текст (возможно, даже на разных языках). В результате чего явного временного соответствия векторов параметров прямым сопоставлением получить не возможно. Ключевым преимуществом данного подхода является удобство и простота настройки системы с таким типом обучения, для конечного пользователя. От него не требуется начитывать определённую базу текстов, можно использовать любые доступные фонограммы или даже аудиозапись голоса в процессе разговора по телефону. Предлагаемый метод текстонезависимого обучения в контексте структуры МГСРТ (рис. 4.)

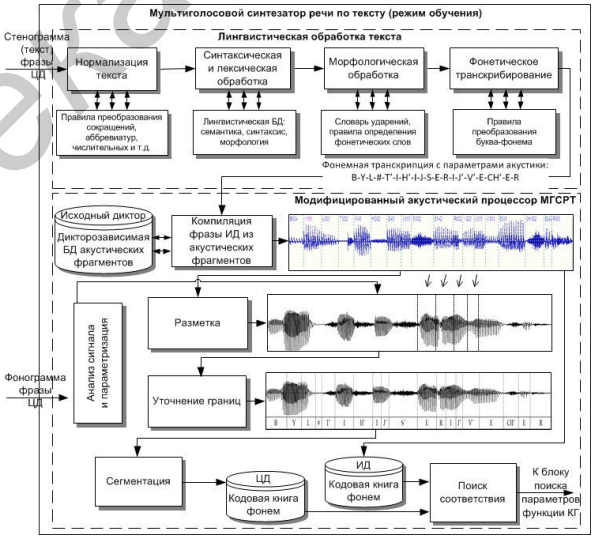

Рис. 1 – Схема текстонезависимого обучения

предполагает активное использование лингвистических блоков СРТ с целью выполнения следующих этапов: нормализации текста, его синтаксической, морфологоческой обработки, фонемного транскрибирования, а также дальнейшей синхронизации полученно фонемной разметки с фонограммой целевого диктора на основе данных полученных от ИД, т.е. синтезатора речи. Этот факт в последствии дает возможность преобразования задачи распознавания фонемных единиц в потоке речи в задачу синхронизации аудиоданных и фонетической разметки и, полученной с текстовых блоков синтезатора ре-

чи. Подробное описание метода можно найти в работе [2].

II. Алгоритм текстонезависимого обучения

В ходе стадии выравнивания параллельно выполняются два процесса: подготовка и анализ речевой и текстовой информации о фонетических и акустических особенностях целевого диктора. Априорной информацией для этих процессов являются последовательность фонограмм обучающих фраз $Wav = (w_1,\ w_2,\ ...,\ w_n)$ и соответствующая ей последовательность орфографической записи этих фраз $Torpho = (t_1,\ t_2,\ ...,\ t_n)$, где n – количество фраз обучающей выборки. Далее над текстовой информацией лингвистическим процессором осуществляется преобразования орфографического текста в фонемный вид L: $Torpho \rightarrow Tphono \in D^{s\times n}$, где s – количество фонетически различимых единиц в одной фразе, а затем фонетический процессор выполняет преобразование фонемного текста, F: $Tphono \rightarrow Tallo \in D^{a\times n}$, где a – количество аллофонов (комбинаций фонем) в одной фразе, в последовательность индексов аллофонов. Необходимо отметить, что алфавит данных индексов совпадает для всех дикторов, поскольку их состав строго определен и неизменен для представителей одной языковой группы. Алгоритмы реализующие данные преобразования подробно изложены в литературе [3].

Блок анализа речевого сигнала выполняет над последовательностью фонограмм W преобразование , основанное на модели сигнала STRAIGHT[4]. Для уменьшения вычислительной сложности последующих этапов и результирующей модели конверсии, сглаженный спектр $P_R(w,t)$ для каждого момента времени заменяется своей огибающей в параметрическом виде с использованием линейных спектральных частот (ЛСЧ). Таким образом, преобразование, выполняемое блоком анализа речевого сигнала, формально можно определить, как $\hat{A}$: $Wav \rightarrow Prm \in D^{p\times m\times n} | \Pr m(m,n) = \{F_0, \Delta F_0, a_1, a_2 ... a_p\}$, где p – кол-во параметров ЛСЧ, m – номер фрейма сигнала, n – номер фонограммы в выборке.

Далее последовательности индексов аллофонов $Tallo$ и векторов параметров сигнала Prm одновременно поступают на входы блока определения границ аллофонов, в котором производится установление оптимального соответствия между ними.

Данная задача эффективно может быть решена с использованием аппарата скрытых марковских моделей (СММ), в рамках которого, последовательность $Tallo$ будет являться последовательностью состояний, а Prm – последовательностью наблюдений. С точки зрения СММ данный процесс заключается в том, чтобы связать оптимальную последовательность состояний с текущей последовательностью наблюдений для модели. Решение данной задачи, формализованной в виде выражения, возможно с помощью использования итерационного алгоритма Витерби.

$$H : (Tallo,\ Prm) \rightarrow (Tallo^{opt},\ Prm),$$

$$\arg\max P(Tallo,\ Prm|\lambda), \lambda \in (A, B, \pi),$$

$$B^{trg} = \{\forall t | Tallo^{opt}(t)\} \Leftrightarrow P^{trg} = \{\forall p | Prm(p)\}$$

где λ – скрытая марковская модель, A – матрица состояний СММ, B – матрица наблюдений СММ, π– матрица начального распределения состояний СММ. Далее путем объединения статистики всех наблюдений, по каждому из состояний, по всем фразам обучающей выборки, возможно найти соответствие векторов параметров, относящихся к определенному аллофону, благодаря равенству аллофонных алфавитов с точки зрения его состава $Ballo^{src}(i) = Ballo^{trg}(i) = Ballo(i) \Rightarrow Prm^{src}(i) \Leftrightarrow Prm^{trg}(i)$. Таким образом, все вышеперечисленные действия в результате выполнения двух этапов, позволяют сформировать совместную последовательность векторов параметров сигнала по фонетическому принципу $Z = (\{Prm^{src}(i),\ Prm^{trg}(i)\}_i), \forall i \in N, i = 1, I,$(где I – количество аллофонов в базе) для его последующего использования в процессе поиска параметров модели конверсии. Более подробно шаги этапа обучения описаны в работах авторов и литературе [5].

Заключение

Предложенный метод и алгоритм имеют следующие преимущества: позволяет осуществлять преобразование входного орфографического текста стенограммы для получения фонемного текста; предоставляет возможность генерации обучающей выборки для исходного диктора на основе фонемного текста целевого диктора; вся необходимая для обучения информация об ИД содержится внутри МГСРТ виде его акустических ресурсов (база акустических эталонов, словарей и правил), либо может быть получена путём синтеза внутри самого МГСРТ по информации от ЦД (последовательность фонем).

1. Stylinau, Y. Voice tracnsformation: A survey / Y. Stylinau. // ICASSP. –2009. –P. 3585–3588.
2. Захарьев, В.А. Текстонезависимое обучение в системе конверсии голоса на базе скрытых марковских моделей и схемы преобразования буква-фонема / В.А. Захарьев, А.А. Петровский // Цифровая обработка сигналов и её приме-нение (DSPA 2013) —Москва: ИПУ РАН. -–С. 327-332.
3. Лобанов, Б.М. Компьютерный синтез и клонирование речи // Минск: Белорусская наука –2008, p. 344.
4. Kawahara H. et al. TANDEM-STRAIGHT: A temporally stable power spectral representation // ICASSP. – 2008. –IEEE. –P. 3933–3936.
5. Захарьев, В.А. Система синтеза речи по тексту с возможностью настройки на голос целевого диктора / В.А. Захарьев, А.А. Петровский, Б.М. Лобанов // Труды СПИИРАН. – Спб: СПИИРАН. – 2014. – №1(32). – С. 82-98.